Plasticity of human replication program during differentiation in relation with change in gene expression and chromatin reorganization

Hanna Julienne

Thèse

en vue de l'obtention du grade de

Docteur de l'Université de Lyon, délivré par l'École Normale Supérieure de Lyon

Discipline : Physique

Laboratoire de Physique de l'ENS Lyon
École doctorale de Physique et Astrophysique de Lyon

présentée et soutenue publiquement le 11/12/2013
par Madame Hanna Julienne

Plasticité du programme spatio-temporel de réplication au cours du développement et de la différenciation cellulaire

Plasticity of the human replication program during differentiation in relation with change in gene expression and chromatin reorganization

---

Directeur de thèse : Alain Arneodo

Après l'avis de :
Marie-Noëlle Prioleau
Bernard Prum

Devant la commission d'examen formée de :
Alain Arneodo, ENS Lyon, Directeur
Benjamin Audit, ENS Lyon, Membre
Emmanuel Barillot, Institut Curie Paris, Membre
Olivier Hyrien, ENS Paris, Membre
David MacAlpine, Duke Institute for Genome Sciences & Policy, Membre
Philippe Pasero, Institut de Génétique Humaine Montpellier, Membre
Marie-Noëlle Prioleau, Institut Jacques Monod Paris, Rapporteur
Bernard Prum, Université d'Évry Val d'Essonne, Rapporteur

Remerciements

Ces trois années de thèse ont été pour moi l'occasion de mûrir tant sur le plan intellectuel que sur le plan personnel. Cette expérience fut riche de sens et j'aimerais témoigner ma gratitude aux personnes qui m'ont accompagnée tout au long de cette épreuve. En premier lieu, je voudrais remercier Alain Arnéodo pour son encadrement rigoureux et son sens du détail. Sa persévérance et son attention continue à l'égard de mon travail m'ont permis de mener à bien cette thèse. Sans cette constance, nombre de projets seraient restés lettre morte. Benjamin Audit a pris part au travail d'encadrement tout au long de la thèse. La rigueur scientifique avec laquelle il m'a conseillée fait de ce travail un travail rigoureux et sûr. Je le remercie aussi de m'avoir fait part de ses idées salvatrices quand je butais sur un problème scientifique.

Je remercie Bernard Prum et Marie-Noëlle Prioleau d'avoir accepté de rapporter ma thèse. Marie-Noëlle Prioleau, biologiste expérimentale, a bien voulu se plonger dans ce travail de statistique multivariée appliquée à la biologie. Je la remercie d'avoir accepté de se confronter à ce domaine et de nous avoir livré son avis de biologiste sans lequel une approche interdisciplinaire n'a de sens. Je remercie Bernard Prum d'avoir relu avec une minutie extrême mon travail et de m'avoir fait part de ses remarques scientifiques alors que j'étais encore en cours de rédaction. Ces réflexions m'ont éclairée et guidée pour l'écriture de mon second chapitre de thèse. Je souhaite remercier les examinateurs de mon jury, Emmanuel Barillot, Olivier Hyrien, David MacAlpine et Philippe Pasero de leur venue et de leurs nombreuses questions pertinentes lors de ma soutenance de thèse. Grâce à eux, ma soutenance fut bien plus qu'une simple formalité et fut extrêmement enrichissante scientifiquement.

Je remercie aussi, Fabien Mongelard et Élise Dumont qui m'ont donné l'occasion d'avoir une expérience positive d'enseignement.

Aussi, je voudrais remercier l'ensemble des personnes chaleureuses et ac-

cueillantes que j'ai pu côtoyer au laboratoire Joliot-Curie et au laboratoire de Physique de l'ENS Lyon. En particulier, j'aimerais remercier Benoît pour les discussions scientifiques que nous avons échangé. Antoine, précédent étudiant d'Alain Arnéodo, m'a beaucoup éclairée sur la pratique scientifique et je reste, à ce jour, admirative de son travail. Je le remercie chaudement pour les nombreuses explications, toujours très claires, dont il m'a fait part durant ma première année de thèse. Je suis pleine de gratitude envers Marianne qui m'a inspirée et m'inspire toujours, a fait germer en moi de nombreuses idées et m'a poussée à ne pas avoir d'a priori sur moi-même. Je suis très heureuse d'avoir partagé mon bureau avec Rasha, Cristina et Laura. Leur joie de vivre et leur chaleur humaine ont été de véritables points d'appuis dans les moments plus difficiles. Je remercie Simona, pour sa passion scientifique et sa gentillesse. Je souhaite bonne chance aux nouvelles arrivantes Guénola et Qiongxu.

Enfin, je remercie ma famille pour son soutien tout au long de ma thèse et surtout mon compagnon de vie, Stephan, à qui je dédie cette thèse.

Contents

Chapter I

Introduction

The initial sequencing of the human genome, a decade ago, was expected to reveal all functional elements encoded in the genomic sequence [1, 2]. Instead, the sequence complexity was much higher than expected making the complete annotation of the human genome, still today, a long way down the road. On the one hand, the human genome sequence put an end to the speculation about the number of protein coding genes, revealing fewer than expected. On the other hand, the number of functional elements was much larger than expected. The quantity of functional elements on a genome was estimated by the fraction of conserved sequences through evolution. Recent comparison with 29 eutherian mammal genomes established that $\sim 4\%$ of the human genome is under purifying selection which exceeds, by far, the content in protein coding sequence ($\sim 1.5\%$). Furthermore, 40% of the functional elements identified have an unknown function [3]. The regulatory amount of DNA is greater than the amount of protein-coding DNA demonstrating that the functional role of the primary DNA sequence is not only to code for proteins but also to regulate nuclear functions including transcription, replication and 3D organization [2,4].

Comparative genomics has the advantage to be independent from the current state of knowledge on the nuclear molecular machinery. Therefore, comparative analyses reveal in an unbiased fashion evolutionarily constrained elements on the genome. However, comparative genomics is unable to assign a role to every discovered elements. By nature, genomics also let unexplored regulatory phenomena that are not encoded in the sequence. Epigenetics emerged as a necessary answer to genomics weaknesses [5, 6]. Epigenomics precisely compensates genomics blind spots by measuring regulatory processes

that let DNA sequences unaltered and by proposing players that could explain the evolutionary constraints. For instance, a CHiP-seq assay can reveal that some functional elements are the binding sites of a transcription factor. Consortium like ENCODE [7] or roadmap epigenomic [8] provide very useful epigenetic databases by describing experimentally the nucleus state in diverse cell types and under a wide range of conditions. In the human genome, of interest here, we have at disposal, gene expression data obtained with RNA-Seq technique [9,10], and genome-wide profiling of Mean-Replication Timing (MRT is the moment of the S-phase at which a locus is replicated, see the precise definition in chapter II) in human [11–14] and in different cell lines. We can also access to chromosomal profiles of many epigenetic modifications [9,15,16], nucleosome positioning [17–19] and chromatin accessibility such as sensitivity to DNase I cleavage [9, 20, 21], that all characterize the primary chromatin structure. In addition, the recent development of the Chromosome Conformation Capture (3C) technology [22], its high-throughput extensions [23–25] including Hi-C [26], and derivatives [27, 28] have provided quantitative measurement of intra- and inter-chromosomal interaction maps [25–30] from which very instructive information can be extracted on the tertiary (3D) chromatin structure and dynamics [26, 29–33]. The different nuclear functions (for instance transcription, replication, 3D organization) are so intertwined that an integrated study seems necessary to assess the interplay between them. All data available, should be, as far as possible, taken into account simultaneously. For instance, differentiation induces important changes in MRT profiles in chromosomal units of size $\sim$ 400-800 kb [34–36]. Early to late (EtoL) MRT changes were associated with loss of pluripotency while late to early (LtoE) changes associated with germ-layer specific transcriptional activation [36]. Importantly, these dynamic changes in MRT come along with some nuclear repositioning [34–40]. EtoL (resp. LtoE) transitions occur simultaneously with a movement from (resp. towards) interior of the nucleus towards (resp. from) a more peripheral location or near nucleoli [40–44]. Transcription is also influenced by MRT changes. Even though most of genes remain at the same expression level in EtoL (reps. LtoE) regions, the number of repressed (activated) genes in these regions is more than expected [34–36]. Additionally, EtoL MRT changes are accompanied by the formation of compact heterochromatin at the nuclear periphery. The four nuclear functions observed (transcription, replication, 3D organization, and chromatin compaction) change successively and impact each other. Therefore, an integrated study seems to be the only

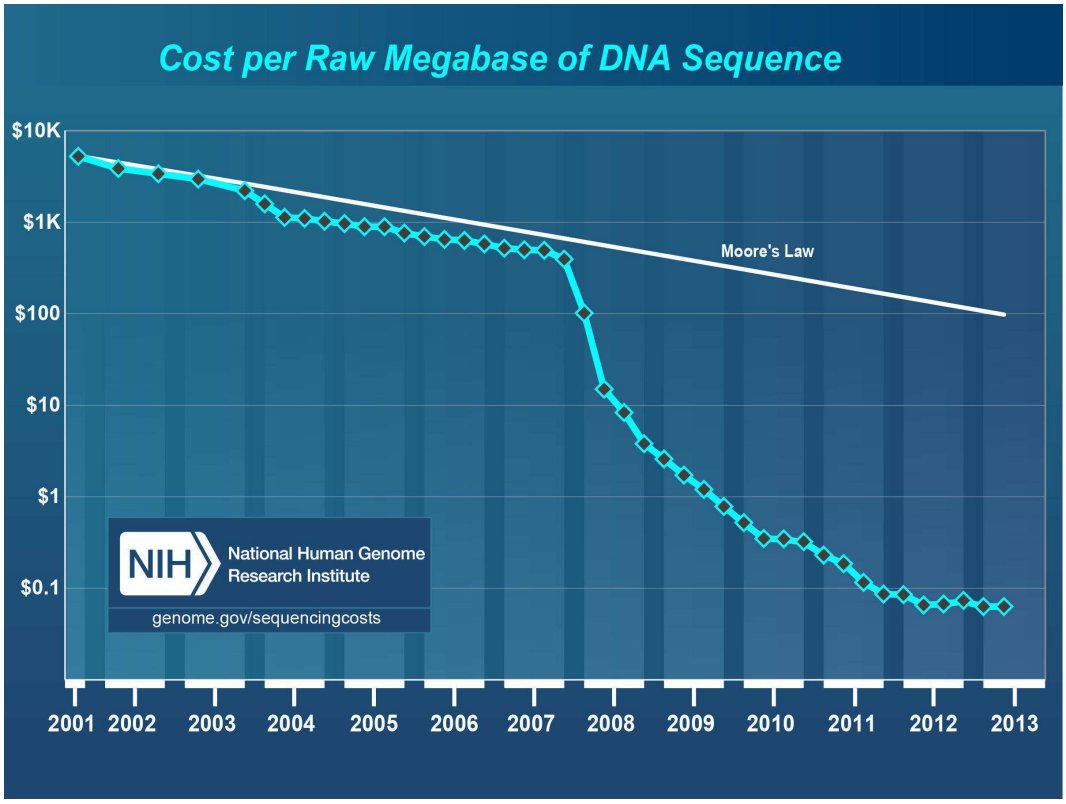

Figure I.1: The cost of DNA sequencing has dramatically fallen during the past twelves years.

way to give the full picture of what is going on in the cell nucleus.

Integrated studies at genome-wide scale are now possible thanks to the high-throughput sequencing technology. Indeed, the falling price of sequencing (Fig. I.1) enables the collection of data on every nuclear process. This trend is not particular to biology, in many fields of application (social network, internet connections, customer transactions) and research (high-throughput sequencing in biology, astronomical data, climate observations), data are intensively collected [45], to the point that if we divided all available data collected between all humans, each person would obtain a quantity of information equivalent to 320 times the library of Alexandria which gathered all human knowledge three century before our era [46]. The promise, inspired by this huge quantity of data, of a better understanding of our surrounding world is accompanied by the daunting task of handling and making sense of these data. In biology, big data require infrastructure and standards to be shared by researchers all over

the world [47]. In this perspective, the ENCODE project is a good direction since it provides standards for both experimental procedure and data formatting and treatment [7, 9, 48, 49]. The second problem with big data in biology is their potential complexity. To illustrate the latent complexity of biological dataset, let us take for example a real integrated study of *Drosophila* chromatin by DamID data analysis [50]. DamID is a technique similar to ChIP-seq which identifies the binding sites of a DNA-associated protein. To characterize the chromatin structure of a genome with all players involved (RNA-binding proteins, chromatin remodelers, histone modifications, histone variants, histone acetyltransferase, histone methyltransferase, *etc*), the authors generated 53 genome-wide DamID profiles. Let us assess the potential complexity of their dataset. To make the argument simpler, we suppose that the DamID signal is binary (the DNA binding protein/histone variant/histone modification is present at one given locus or absent). To study comprehensively the chromatin structure, we have to look at all possible combinations. A simple calculation leads to the conclusion that, for each locus, we would have to consider 2^{53} combinations *i.e.* 9007199 billion of cases. In other words, looking at combinations one by one is impossible. Fortunately, chromatin does not explore all possible combinations. Indeed, fluorescence assays show that some proteins colocalize in the nucleus whereas others segregate. Statistically speaking, it means that a lot of information contained in these signals is redundant. Statistical analyses, taking advantage of this redundancy, have shown that this huge combinatorial complexity can be reduced to a surprisingly small number of predominant chromatin states with shared features namely four in *Arabidopsis thaliana* [51], five in *Caenorhabditis elegans* [52] and four [53] or five in our example [50] in *Drosophila*. Fig. I.2 shows how successful this analysis was, starting from 53 intertwined, unreadable DamID profiles, they ended with 5 distinct chromatin states with an easy interpretation for each one of them.

In this thesis, our focus will be placed on DNA replication. DNA replication, the basis of genetic inheritance, is of fundamental importance to cellular life: when a cell fails to regulate its replication program, it strongly affects the genome integrity, which can lead to cell death or cancer. We would like to shed a new light on human DNA replication by taking advantage of the huge set of data available on human chromatin primary structure. How epigenetic mechanisms and gene expression coordinate with DNA replication has been a long-standing question [4,6,54–57]. Contrary to bacteria, yeast and viruses, the genomes of multi-cellular eukaryotes have no clear consensus DNA sequence

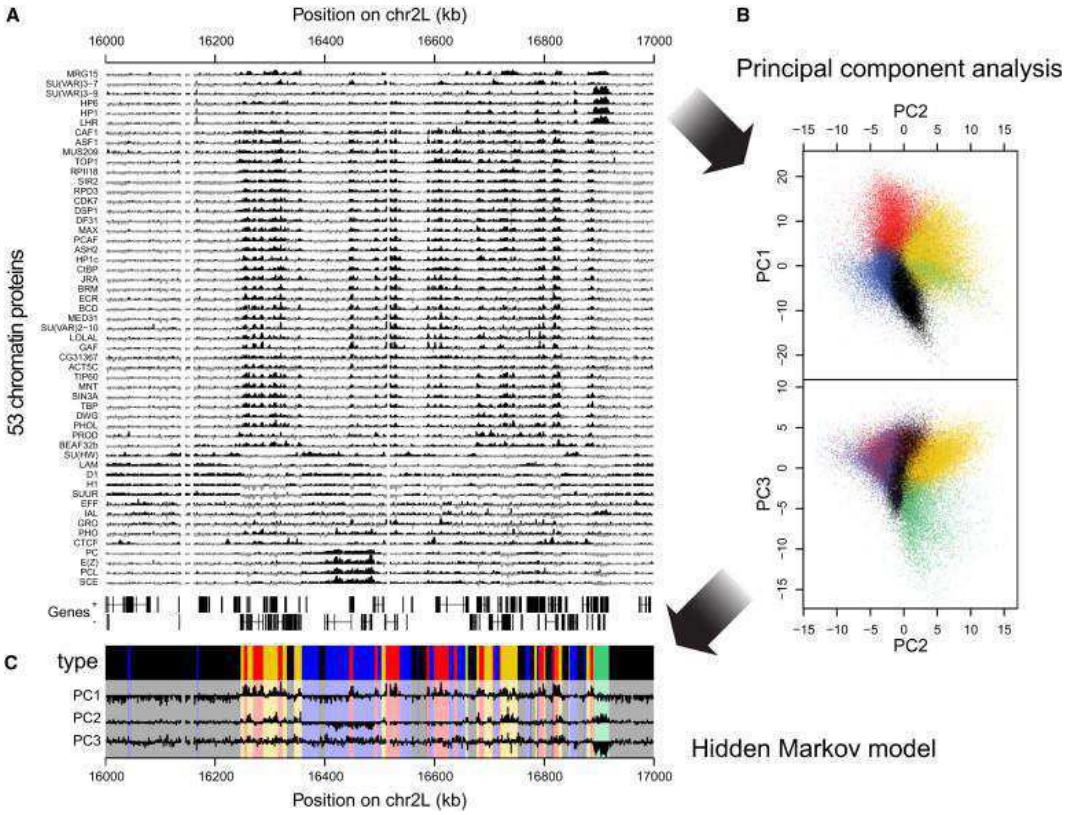

Figure I.2: Example of an efficient dimensionality reduction applied to 53 DamID profiles. (A) Sample plot of all 53 DamID profiles (log_2 enrichment over Dam-only control). Below the profiles, genes on both strands are depicted as lines with blocks indicating exons. (B) Two-dimensional projections of the data onto the first three principal components. Colored dots indicate the chromatin type of probed loci as inferred by a five-state Hidden Markov Model (HMM). (C) Values of the first three principal components along the region shown in (A). Domains of the different chromatin types are highlighted by the same colors as in (B). Reproduced from [50].

element associated with replication initiation [58, 59]. Metazoan genomes duplicate through the coordinated activation of hundreds to thousands of replication origins that can be extremely site-specific or poorly defined with a broad site specification [60]. Indeed more origins are prepared in G1-phase than actively needed in S-phase [61]. Epigenetic mechanisms very likely take part in the spatial and temporal control of origin usage and efficiency in relation with gene expression [61–67]. For many years, elucidating the determinants that specify replication origins has been hampered by the very limited number of well established origins in human and more generally in mammals (a few

tens versus a few ten thousands expected) [57, 61, 66, 68]. Only very recently, nascent DNA strands synthetized at origins were purified by various methods to map replication origins genome-wide in different eukaryotic organisms including *Arabidopsis thaliana* [69], Drosophila [70], mouse [70, 71] and human [16, 72–77]. Despite some inconsistency or poor concordance between certain of these studies [57, 78], some general trends have emerged confirming the correlation of origin specification with transcriptional organization [56, 57, 61]. The set of replication origins identified so far are strongly associated with annotated promoters and seem to be enriched in transcription factor binding sites [73, 74, 79] and in CpG islands [70, 71, 73]. However a significant proportion of origins do not seem to be controlled in the same way as gene transcription since they are in regions void of DNase-I-hypersensitive sites (DHSs) and of histone marks found at active promoters [56, 73]. Interestingly, it has been recently reported that replication origins may contain specific nucleotide sequences. Actually G-rich consensus motifs were shown to be associated with *Drosophila*, mouse and human origins [70, 77, 80]. These analyses have opened new perspectives towards the identification of mechanisms governing origin selection in mammals.

The recent blooming of genome-wide mean-replication timing (MRT) data in yeast [81], plants [82], worm [83], fly [84, 85], mouse [34, 35, 86] and human [11–14] has provided the opportunity to establish links between the spatio-temporal program of replication, transcription and chromatin structure [4,6,56, 57,87]. It is now well established that in higher eukaryotes, GC-rich, gene-rich and early replicating regions colocalize with, as a counterpart, a colocalization of AT-rich, gene poor and late replicating regions [11, 86, 87]. But recent studies in mammals [12, 34] and *Drosophila* [88], have shown that during differentiation, some genes change expression without changes in MRT and vice versa, thereby indicating that transcription is not the only controlling factor and that the chromatin structure is likely to be part of the game [6, 56, 57, 87]. In good agreement with previous studies in *Drosophila* [50, 88], genome-wide MRT profiles along mouse and human chromosomes in different cell lines reveal a correlation with epigenetic modifications [89]. Early replicating regions tend to be enriched in open chromatin marks H3K4me1, H3K4me2, H3K4me3, H3K36me3, H4K20me1 and H3K9 and H3K27 acetylation, whereas late replicating zones are mostly associated with H3K9me2 and to a lesser extent with H3K9me3 [34, 36, 62]. The dynamic changes in MRT observed during development come along with some subnuclear repositioning [34–40], early replicating

euchromatin domains being generally at the interior of the nucleus whereas late replicating heterochromatic domains are more peripheral or near nucleoli [40–44]. Recent experimental studies of long-range chromatin interactions using chromosome conformation capture techniques [26, 30, 36, 90] have confirmed that 3D chromatin tertiary structure plays an important role in regulating replication timing. In particular, replicon size, which is dictated by the spacing between active origins, correlates with the length of chromatin loops [67, 91, 92]. But as questioned in Refs [30, 93, 94], the dichotomic picture proposed in early studies [26,36,90], where early and late replicating loci occur in separated compartments of open and closed chromatin respectively, is somehow too simple as previously pointed out in a detailed analysis of replication fork velocity [93]. Identifying the chromatin regulators of the spatio-temporal program of DNA replication will be a formidable step towards understanding the so-called replicon and replication foci [43, 95–98] in relation with their transcription counterpart, the transcription factories [43, 98–100].

In a recent work [94, 101], the analysis of genome-wide MRT data in seven human cell types including Embryonic Stem Cell (ESC), somatic and HeLa cells, revealed that, in each cell type, about half of the genome can be paved by the so-called replication U-domains where the MRT is U-shaped and its derivative N-shaped like the nucleotide compositional asymmetry in the germline skew N-domains [102–106]. These N-shaped patterns are the consequence of large-scale gradients of replication fork polarity [94, 107, 108] originating from early initiation zones separated by several megabases. In that regard, N/U-domains can be thought as an equivalent of bacterial replicon [109]. These "master" replication origins [4, 110] at U/N-domain borders were found to be hypersensitive to DNase I cleavage, to be transcriptionally active and to display a significant enrichment in the insulator binding protein CTCF, the hallmarks of localized ($\sim$ 200-300 kb) open chromatin structures [94, 111, 112]. A cascade model of origin firing was recently proposed to account for the observed progressive inversion of replication fork polarity inside U/N-domains [93, 109]. This model involves the superposition of specific and efficient initiations at domain borders with random and less efficient initiations elsewhere, in addition to firing stimulated by propagating forks. The comparative analysis of chromatin interaction Hi-C [94] and 4C [30] data with MRT profiles further confirmed that these replication U/N-domains actually correspond to topological domains of self-interacting chromatin. As recently demonstrated using a graph theoretical approach [113], master replication origins at U/N-domain borders

are long-range interconnected hubs of chromatin interactions both within and between different chromosomes. The additional observation of a remarkable gene organization inside U/N-domains with a significant enrichment of expressed genes nearby their borders [94, 104, 114] prompted the interpretation of these replication domains as chromatin units of highly coordinated regulation of transcription and replication [94, 109, 110]. Replication U/N-domains are also likely to be central to genome regulation since the dynamical changes in MRT profiles observed during differentiation [34–36, 115] mainly occur in the 50% of the genome that are covered by U/N-domains [94]. Overall, these results point out that U/N-domain borders offer a good framework to the understanding of the plasticity of the spatio-temporal replication program, gene expression and chromatin organization across different cell lines during development and lineage commitment.

In this thesis, we conduct an integrative analysis of the interplay between the chromatin primary structure and the MRT. To do so, we simplify a dataset of several genome-wide CHiP-seq profiles to four prevalent chromatin states that have strongly different MRT distributions. We use U/N-domains as a guide to study the spatial distribution of these chromatin states with respect to the spatio-temporal replication program. The genome-wide CHiP-seq data allow us to assess the distribution of these chromatin states in the 50% of the human genome not covered by U/N-domains. U/N-domains are also our framework to describe coordinated changes of chromatin composition and replication program through development.

The manuscript is organized in six chapters. The first chapter is the current introduction. Chapter II is devoted to definitions and methodological discussions that explain biological (eukaryotic DNA replication, mean replication timing, U-domains, chromatin primary structure and its potential causal link to replication) and statistical (Spearman correlation, Principal component analysis, clustering) concepts used in the "results" chapters. The results are reported in Chapters III, IV and V. Chapter III deals with an integrative analysis of the genome-wide distributions of thirteen epigenetic marks, at 100 kb resolution, in the human cell line K562. This integrative analysis identifies four major groups of chromatin marks with distinct features. These chromatin states have different MRT, namely from early to late replicating, replication proceeds through a transcriptionally active euchromatin state (C1), a repressive type of chromatin (C2) associated with polycomb complexes, a silent state

(C3) not enriched in any available marks, and a gene poor HP1-associated heterochromatin state (C4). Chapter IV is a reproduction of our integrative analysis of epigenetic data in the K562 human cell line at a smaller scale (6 kb) characteristic of gene promoters. By investigating the coherence between the chromatin states obtained at 100 kb and 6 kb, we will assess to what extent promoter activity conditions its large-scale chromatin environment and *vice-versa*. In Chapter V, we extend this study to different cell types including the ESC H1hesc, three hematopoietic cell lines (K562, Gm1278, Monocyte CD14+), a mammary epithelial cell line (Hmec) and an adult fibroblast cell line (Nhdfad). By exploring the global reorganization of replication U/N-domains in these different cell types in relation to coordinated changes in chromatin state and gene expression, we shed a new light on the chromatin-mediated epigenetic regulation of transcription and replication during differentiation. Because they are likely to be the cornerstone to a better understanding of pluripotency maintenance and lineage commitment, we will pay special attention to the "master" replication initiation zones that border U/N-domains and specially to those that are specific to ESCs. Chapter VI is a general discussion that deals with the interpretation of the reported results. Do the results, gathered in this thesis, give any information on the causality between primary chromatin structure and replication? Some perspectives will be also given on how the methodology described in Chapters III, IV and V could be applied to cancer.

Chapter II

Definitions and methodology

The current chapter explains the biological concepts and statistical methodology necessary to understand the chapters reporting results and the discussion. The first section presents basic prerequisites on DNA replication, and gives the definition of the Mean Replication Timing (MRT) and of the U-domains. The second section discusses chromatin structure and plausible causal links between the primary structure of chromatin and DNA replication. The third section is a brief presentation of the experimental protocols used to produce datasets analyzed in this thesis. The fourth part of the chapter is an educational presentation, on simple artificial examples, of the tools (principal component analysis and clustering) used in the "results" chapters.

II.1 Prerequisites on DNA replication

II.1.1 Eukaryotic cell cycle and DNA replication

Living cells, prokaryotes[1] and eukaryotes[2], use a universal mechanism to multiply. Cells duplicate their content, among which DNA, and divide. In eukaryotic cells, this phenomenon is decomposed into 4 phases which form the cell cycle (Fig. II.1). Cell division occurs during the mitosis or M phase. The rest of the cell cycle, called the interphase, is subdivided in the G1, S, and G2

[1]The prokaryotes are a group of organisms (bacteria for example) whose cells lack a membrane-bound nucleus (karyon).

[2]An eukaryote is an organism (yeast or human for example) whose cells contain a nucleus and other structures (organelles) enclosed within membranes.

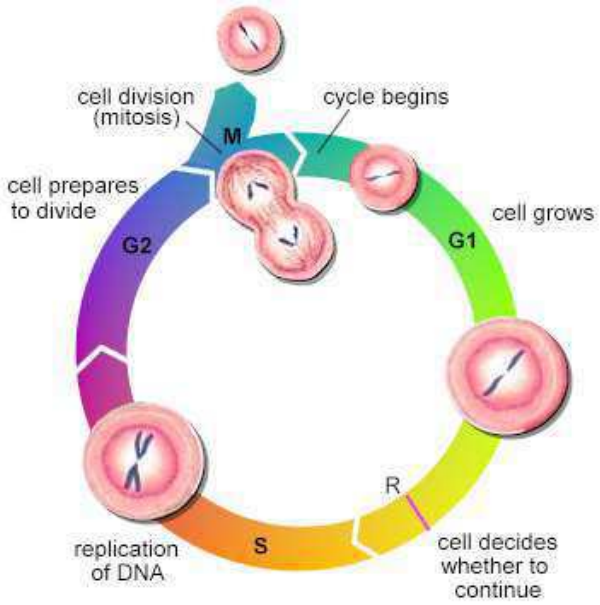

Figure II.1: The four phases of an eukaryotic cell cycle. The union of G1, S and G2 is called the interphase, during which the cell grows continuously. The cell divides in two new cells in M phase. The S phase, of particular interest here, is the phase where DNA is replicated.

phases. Cells grow continuously during the interphase, doubling in size and preparing for the next division. Yet, DNA duplication occurs only during the S phase.

DNA replication is an essential genomic function responsible for the accurate transmission of genetic information through successive cell generations. Replication follows a simple pattern in bacteria called the "replicon" model [116]; the process starts by the binding of some "initiator" protein complex to a consensus "replicator" DNA sequence called origin of replication. On a bacterial genome, there is only one origin of replication. The recruitment of additional factors initiates the bi-directional progression of two divergent replication forks along the chromosome. Since most bacterial genomes are circular, the two forks join at the terminus of replication usually located at the opposite of the origin on the circular chromosome. One strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the origin (lagging strand) (Fig. II.2B).

In eukaryotic cells, replication is initiated at a number of replication origins (more than 30000 in the human genome) and propagates until two converging forks collide at a terminus of replication [59,117]. During G1 phase, the Origin

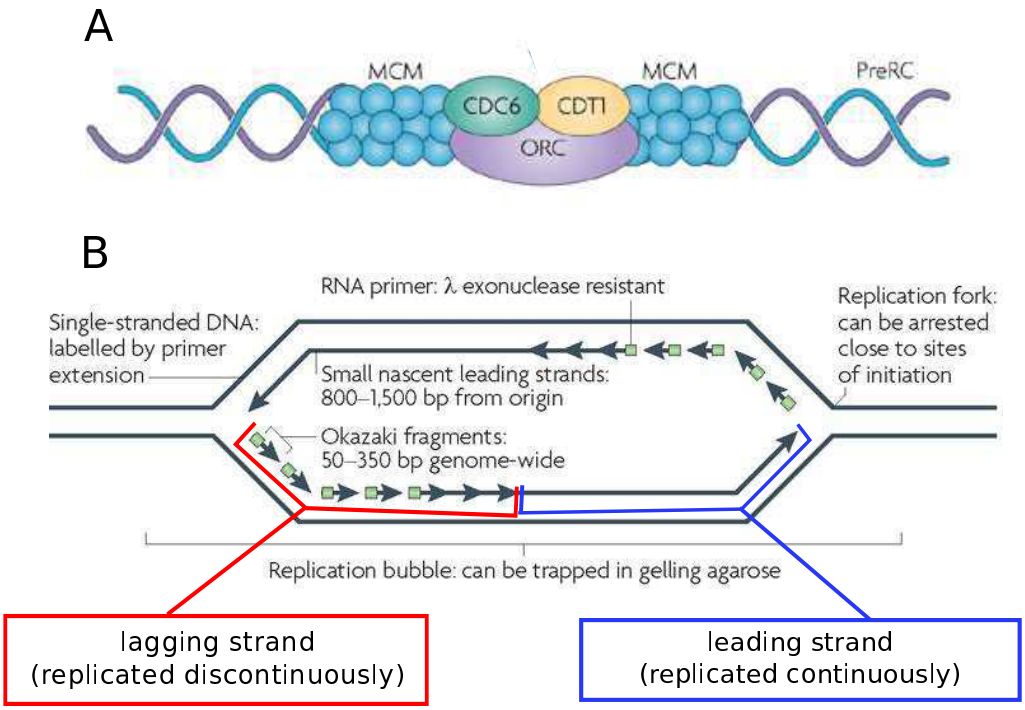

Figure II.2: (A) The six-subunit complex called the origin recognition complex (ORC) serves as a platform for the assembly of pre-replication complexes. In metazoans, the binding of the large subunit of ORC, ORC1, to chromatin is cell-cycle regulated. During the mitosis/G1-phase transition, chromatin-bound ORC recruits CDC6 and CDT1, which facilitate the loading of a helicase complex consisting of 2 to 7 MCM (minichromosome maintenance) proteins. The resulting complex is termed the pre-replication complex (PreRC). Reproduced from [66]. (B) Replication bubble structure and summary of the unique nucleic acid features found near origins of replication. The leading strands of DNA synthesis quickly become larger than Okazaki fragments and can be isolated as small single-stranded molecules that can be verified to be nascent either by metabolic labeling or by virtue of the fact that nascent strands have small stretches of RNA at their 5′ ends that render them resistant to λ-exonuclease. The topological structure of replication origins shortly after initiation is a bubble structure, which can be trapped in gelling agarose. Reproduced from [57].

Recognition Complex (ORC) binds to DNA. The binding of ORC is followed by the recruitment of several proteins including the helicase MCM (Minichromosome maintenance). These proteins form the pre-replication complex (pre-RC) pictured in Fig. II.2A. The pre-RC constitutes a potential replication origin that may be activated during S-phase. In fact, there are more pre-RC deposited on DNA than actively needed during the S-phase. The subsequent activation of the pre-RC during S-phase leads to the recruitment of DNA polymerase and other proteins necessary to the DNA synthesis. The activation of different replication origins occurs at diverse moments of the S phase and is not deterministic [84, 109, 118–121]. The pre-RC activation can be triggered by the activation of neighboring replication origins [93]. Also, the activation depends on the neighboring transcriptional activity and on the local chromatin structure [84, 119–121].

Sequence requirements for a replication origin vary significantly between different eukaryotic organisms. In the unicellular eukaryote *S. cerevisiae*, the replication origins spread over 100-150 bp and present some highly conserved motifs [59]. However, the prokaryote-like replication of *S. cerevisiae* is an exception among eukaryotes. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the replication origins spread over at least 800 to 1000 bp [59]. In multicellular organisms, the nature of initiation sites of DNA replication is even more complex [117]. Metazoan[3] replication origins are rather poorly defined and initiation may occur at multiple sites, each site being distributed over a thousand of base pairs [60, 109]. The initiation of replication at random and closely spaced sites was repeatedly observed in *Drosophila* and *Xenopus* early embryo cells, presumably to allow for extremely rapid S phase, suggesting that any DNA sequence can function as a replicator [109, 118, 122, 123]. A developmental change occurs around midblastula[4] transition that coincides with some remodeling of the chromatin structure, transcription ability and selection of preferential initiation sites [118, 123]. Thus, although it is clear that some sites consistently act as replication origins in most eukaryotic cells, the mechanisms that select these sites and the sequences that determine their location remain elusive in many cell types [63, 117, 124]. As recently proposed by many authors [64, 65, 125],

[3]Metazoan are multicellular organisms belonging to the kingdom Animalia (according to Linnaeus' classification).

[4]The blastula is a hollow sphere of cells formed during an early stage of embryonic development in animals.

the need to fulfill specific requirements that result from cell diversification may have led high eukaryotes to develop various epigenetic controls over the replication origin selection rather than to conserve specific replication sequence.

This might explain that for many years, very few replication origins have been identified in multicellular eukaryotes, namely around 20 in metazoa and only about 10 in human. Several techniques have been used to detect replication origins. A first technique takes advantage of the presence of the ORC proteins at the origins to detect their position by CHiP-seq[5] [126–128]. Alternatively, in recent studies, nascent DNA strands synthesized at origins were purified by various methods to map replication origins genome-wide in different eukaryotic organisms including *Arabidopsis thaliana* [69], Drosophila [70], mouse [35, 70, 71] and human [11–14, 16, 72–77]. Another approach to discover replication origins is to trap replication bubbles [75, 129]. The DNA particularities around replication origins that have been used to position origins by these different techniques are summarized in Fig. II.2B. Despite some inconsistencies or poor concordance between certain of these studies [57, 78], some general trends have emerged confirming the correlation of origin specification with transcriptional organization [56, 57, 61]. The set of replication origins identified so far are strongly associated with annotated promoters and seem to be enriched in transcription factor binding sites [73, 74, 79] and in CpG islands [70, 71, 73].

An alternative to characterize replication is to estimate at what moment of the S-phase a locus is replicated (replication timing). A wealth of genome-wide replication timing data is available for several eukaryotic organisms ranging from yeast [81], to plants [82], to worm [83], to drosophila [84], to mouse [34,35], and to human [11–14, 130]. Recent genome-wide replication timing data has been collected in several human cell types [11–14, 36, 90, 130], which enables to study changes in the replication program across differentiation. In this thesis, we focus our study on the current abundance of replication timing data. The replication timing at a given locus depends on the local initiation properties, but it also depends on the initiation properties of neighboring sites as replication forks propagate [131, 132]. Therefore replication timing can be difficult to interpret.

To clarify the definitions of replication timing and spatio-temporal replication program, we now describe an idealized example for one cell cycle.

[5]The CHiP-seq protocol is briefly presented in Sect. II.3.1.

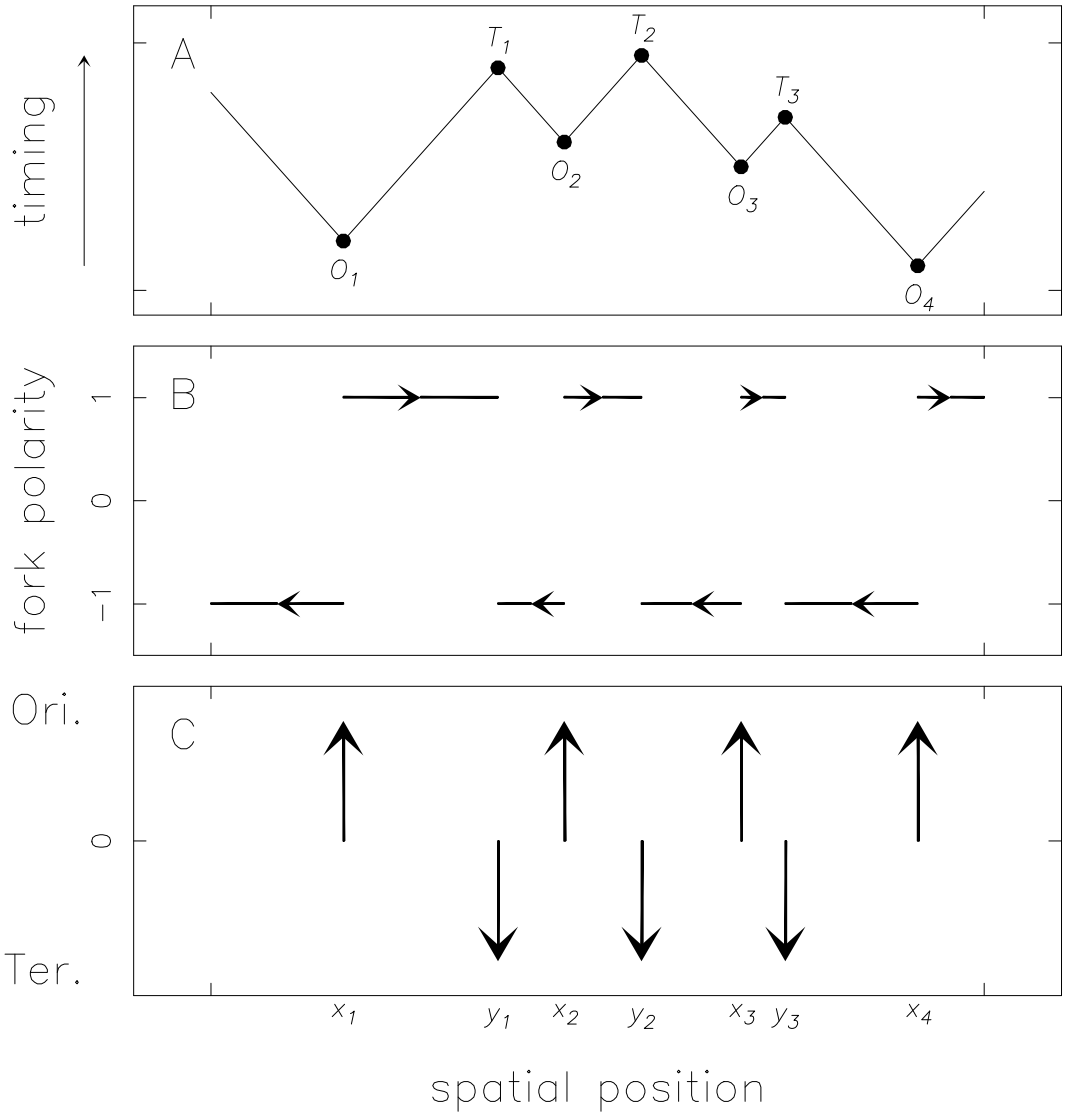

Figure II.3: Replication program in one cell cycle. (A) Replication timing $t_R(x)$, (B) replication fork orientation $o(x)$ and (C) spatial location of replication origins (upward arrows) and termination sites (downward arrows). $O_i = (x_i, t_i)$ corresponds to the origin i positioned at location x_i and firing at time t_i. Fork coming form O_i meets the fork coming from O_{i+1} at termination site T_i with space-time coordinates (y_i, u_i) given in Eq. (II.1). Note that we can deduce the fork orientation in (B) (resp. origin and termination site locations in (C)) by simply taking successive derivatives of the timing profile in (A) (Eqs. (II.3) and (II.4)).

II.1.2 Replication program in one cell cycle

If the replication fork velocity v is constant, the spatio-temporal program of replication for one cell cycle is completely specified by the positions x_i and the

firing times t_i of the n activated bidirectional replication origins O_i (Fig. II.3A). From each bidirectional origin, two divergent forks propagate at velocity v, until they meet a fork of the opposite orientation. Let T_i be the termination locus where the fork coming from O_i meets the fork coming from O_{i+1}. Straightforward calculations lead to the space-time coordinates (y_i, u_i) for T_i:

$$y_i = \frac{1}{2}(x_{i+1} + x_i) + \frac{v}{2}(t_{i+1} - t_i), \quad u_i = \frac{1}{2v}(x_{i+1} - x_i) + \frac{1}{2}(t_{i+1} + t_i). \quad \text{(II.1)}$$

In Fig. II.3, the x-axis is conventionally oriented in the $5' \to 3'$ direction of the reference strand. Hence sense $(+)$ and antisense $(-)$ forks, correspond respectively to rightward and leftward moving forks in Fig. II.3B.

Around origin O_i (for $x \in [y_{i-1}, y_i]$), the replication timing $t_R(x)$ and the fork orientation $o(x) = \pm 1$ are given by:

$$t_R(x) = t_i + |x - x_i|/v \quad \text{and} \quad o(x) = \text{sign}(x - x_i). \quad \text{(II.2)}$$

Finally, using the Dirac distribution δ to represent origin locations $\delta(x - x_i)$ and termination sites $\delta(y - y_i)$ (Fig. II.3C), we obtain the following fundamental relationships:

$$v \, \frac{d}{dx} t_R(x) = o(x), \quad \text{(II.3)}$$

$$v \, \frac{d^2}{dx^2} t_R(x) = \sum_i \delta(x - x_i) - \sum_i \delta(x - y_i). \quad \text{(II.4)}$$

In other words, we can extract, up to a multiplicative constant, the fork orientation $o(x)$ (Fig. II.3B) and the location of origin and termination sites (Fig. II.3C) by simply taking successive derivatives (Eqs. (II.3) and (II.4)) of the timing profile $t_R(x)$ (Fig. II.3A).

II.1.3 Mean Replication Timing

This paragraph describes how we have extracted Mean Replication Timing (MRT) [94] from experimental data provided in [14].

Current technology is not able to measure the spatio-temporal replication program in one cell. The characterization of replication timing is done on a large population of cells (a few millions). Briefly, the method used to generate experimental data in [14] is as follows:

* A large population of cells, each at a different moment of the cell cycle, is temporarily cultivated in presence of BrdU which is a modified nucleotide. Cells in S-phase incorporate BrdU in place of thymidine in newly synthesized DNA which is, hence, identifiable.

* Cells are classified according to their DNA content by Fluorescence-Activated Cell Sorting (FACS) which is equivalent to classify them according to their cell cycle phase. Indeed, cells start with one copy of their genome in G1-phase, they gradually double their DNA content through S-phase and have two genome copies in G2. Then they undergo mitosis that sets their DNA content back to one copy. The classification by FACS has a limited time resolution. In the study of interest [14], they dispose of 6 bin spanning the S-phase (Fig II.4A).

* Once the cell population is classified into 6 bins (G1b, S1, S2, S3, S4, G2), the newly synthesized DNA (*i.e.* DNA that contains BrdU nucleotides) is sequenced and mapped on the genome. For each temporal bin, the density of tags is computed genome wide (Fig II.4B).

To efficiently summarize the information contained in the six temporal bin, we applied the following post-treatment in [94] :

* A value is assigned to each temporal bin. By convention, the very beginning of S-phase is zero and the very end is 1. Since the temporal resolution is of 6 bins, each bin are one sixth long, the values attributed are 1/12 for G1b, 3/12 for S1, 5/12 for S2, 7/12 for S3, 9/12 for S4 and 11/12 for G2.

* In 100 kb sliding windows incremented by 10kb steps, tags are retrieved for each of the six temporal bins.

* The mean replication timing for one given window is the sum of the proportions of tags found in each bin multiplied by the corresponding bin timing.

For instance, if in a fairly early 100 kb window the read proportion in the six bins is as follows: 20% , 50%, 20%, 7%, 2%, 1%, the MRT is:

$$\text{MRT} = 0.2 \times \frac{1}{12} + 0.5 \times \frac{3}{12} + 0.2 \times \frac{5}{12} + 0.07 \times \frac{7}{12} + 0.02 \times \frac{9}{12} + 0.01 \times \frac{11}{12} = 0.29$$

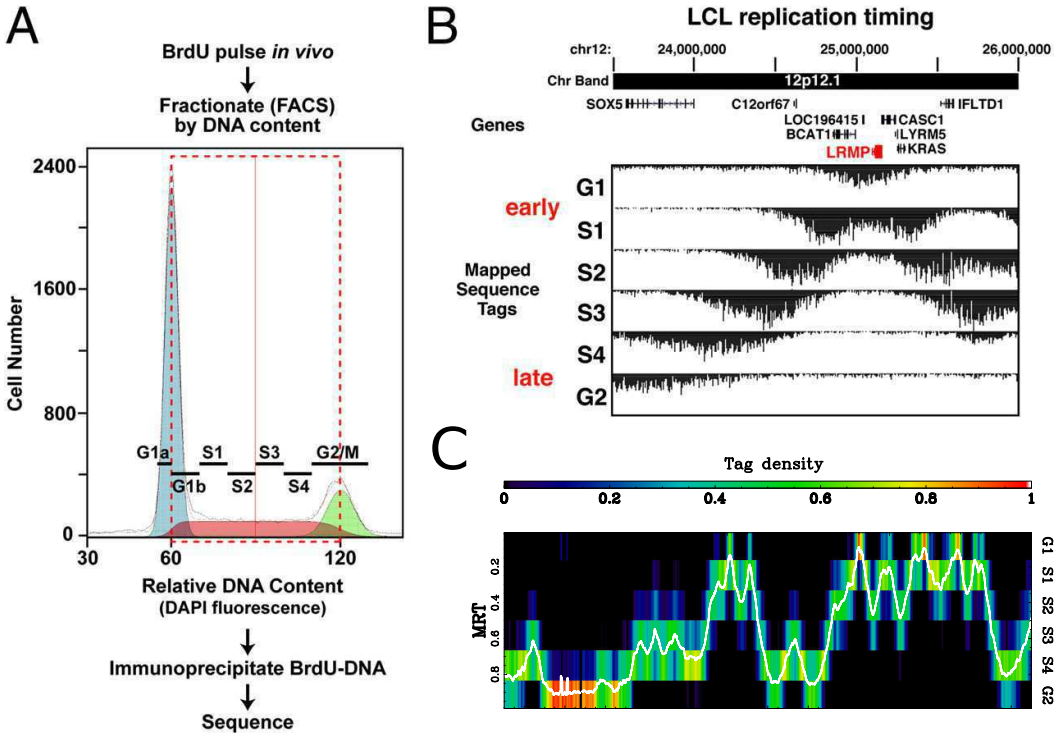

Figure II.4: (A) Classification of the cell population according to their DNA content. The fractions retained for MRT computation are those marked with G1b, S1, S2, S3, S4, and G2 labels. In each fraction, the DNA containing BrdU is retrieved by immunoprecipitation. Then, the DNA is sequenced. (B) Brut sequence tag density on the LMRP locus in GM06690. (A) and (B) are reproduced from [14]. (C) Post treatment applied in [94] to compute MRT. Normalized tag densities on a 25 Mb long fragment of chromosome 10 for the GM06990 cell line, and the corresponding computed MRT (white line).

MRT interpretation

Even though the replication timing in each cell is a random variable, the MRT is a good indication of the replication timing population distribution. Indeed, BrdU reads are generally located in few temporal bins at one locus [94]. This suggests that the replication spatio-temporal program is fairly constant in cells of the same cell line. Suppose we observe a MRT = 0.5. Theoretically, it could mean that the replication timing distribution is bimodal with one half of the cell population replicating early and the other half late. Alternatively, it could mean that the replication timing distribution is uniform (*i.e* the chances of being replicated in any of the six bin are equal). In practice, a MRT of 0.5 means that, in the vast majority of cells, the locus is replicated in mid-S phase.

II.1.4 Replication U-domains

Replication is an asymmetrical process. Indeed, one strand is replicated continuously (leading strand) while the other strand is replicated in a step like fashion (Fig. II.2B). This phenomenon imprints the DNA sequence through evolution: the mutation rates are different on the leading and on the lagging strands. This discrepancy induces a compositional skew $S = \frac{T-A}{T+A} + \frac{G-C}{G+C}$ [4, 102–107] that reflects the replication fork polarity [94, 107].

Compositional asymmetry in bacteria

A clear relationship between replication and compositional asymmetry was first established in prokaryotic genomes by Lobry [134]. In bacteria, the spatio-temporal replication program is particularly simple (Fig. II.5A). The replication origin is defined by a consensus sequence, replication therefore always initiates at the same genomic locus (ORI), two divergent forks then replicate the DNA until they meet at the replication terminus (TER). As shown in Fig. II.5B for *Bacillus subtilis*, many prokaryotic genomes are divided into two halves: one presents an excess of guanine over cytosine, and the other one, on the opposite, an excess of cytosine over guanine. The GC skew, defined as $S_{GC} = \frac{G-C}{G+C}$, is thus positive on one half of the genome and negative on the other. Remarkably, the GC skew profile is tightly related to the spatio-temporal replication program: the leading strand has positive GC skew whereas the lagging strand has negative GC skew.

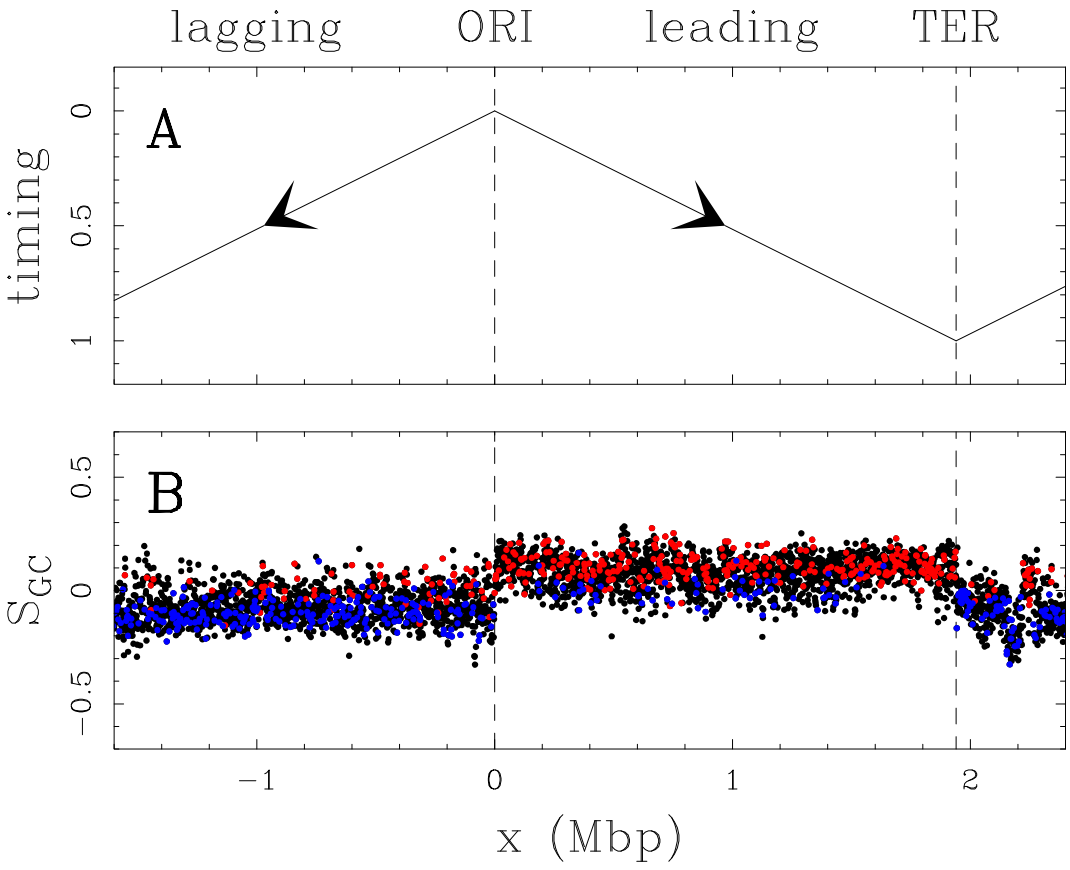

Figure II.5: Comparing GC skew $S_{GC} = \frac{G-C}{G+C}$ and replication timing in *Bacillus subtilis* genome. (A) Schematic representation of the replicon model: divergent bidirectional progression of the two replication forks from the replication origin (ORI) to the replication terminus (TER). The replication timing is indicated from early, 0 to late, 1. (B) S_{GC} calculated in 1 kbp windows along the genomic sequence of *Bacillus subtilis*. Black points correspond to intergenic regions, red (resp. blue) points correspond to (+) (resp. (−)) genes, which coding sequences are on the published (resp. complementary) strand. Reproduced from [133].

Compositional asymmetry in the human genome

By contrast, the spatio-temporal replication program in eukaryotes is much more complex. Several initiation sites are used in each cell cycle, and they fire at different times during the S phase. Furthermore, the genomic positions and firing times of the initiation sites change from one cell cycle to another. Yet, the relationship observed between the compositional asymmetry and the replication program in bacteria can be generalized to eukaryotic genomes.

In the human genome, the skew profile presents N-shaped domains of sev-

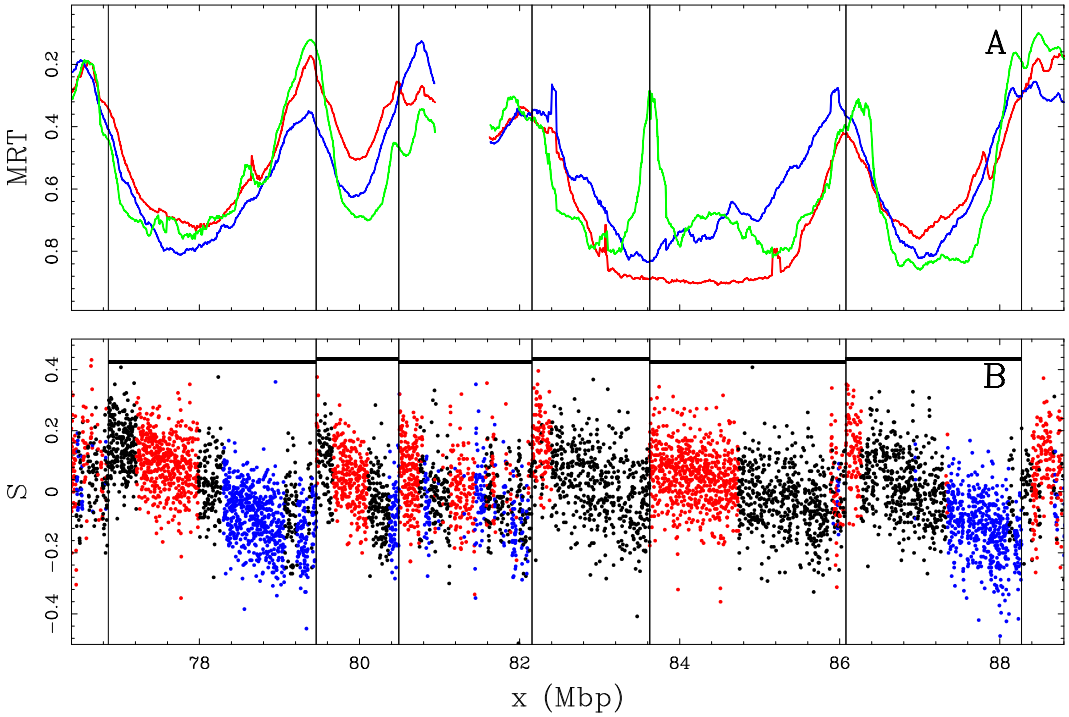

Figure II.6: Comparing compositional skew $S = \frac{T-A}{T+A} + \frac{G-C}{G+C}$ and mean replication timing (MRT) in the human genome. (A) MRT profiles along a 11.4 Mbp long fragment of human chromosome 10, from early (0) to late (1) for BG02 embryonic stem cell (green), K562 erythroid (red) and GM06990 lymphoblastoid (blue) cell lines. Replication timing data was retrieved from [14]. (B) S calculated in 1 kbp windows of repeat-masked sequence. The colors correspond to intergenic (black), (+) sense genes (red) and (−) antisense genes (blue). Six skew N-domains (horizontal black bars) were detected in this genomic region. Reproduced from [133].

eral megabases. Previous works have led to the objective delineation of these N-shaped skew domains [102–105,135]. Based on the analogy with the bacterial case (the upward jump of the GC skew colocalizes with the ORI in Fig. II.5), the N-domains borders (upward jumps of the skew) were proposed to be replication origins, evolutionary conserved and active in the germline [102–104]. However, we know today that the N-shape skew profile is not a trivial extension of the replicon model in bacteria. For instance, the typical inter-origin distance ($\sim$ 40 kb) measured using DNA combing [93], is much smaller that the typical N-domain size ($1-3$ Mb), which implies that many other initiation events occur inside the N-domains. A recent theoretical analysis [107] demonstrates that the skew profile, in bacteria and in the human, actually results

from the replication fork polarity profile. In fact, the fork polarity follows the same trend as the skew in N-domains going from a positive value at the N-domain left border and decreasing to negative opposite value at the N-domains right border.

We observe in Fig. II.6 a clear relationship between the compositional asymmetry and the replication timing in the human genome: a N-shaped compositional skew $S = \frac{G-C}{G+C} + \frac{T-A}{T+A}$ profile remarkably corresponds to a U-shaped replication timing profile. A previous analysis [94] demonstrates that the derivative of the replication timing is the replication fork polarity (Fig II.3). Importantly, this relation remains true for a population of cells. Mathematically, by taking the derivative of a U-shaped function, a N-shaped pattern is obtained. Therefore the mean polarity of the replication forks is a N-shaped function in U-domains (Fig II.7), confirming the existence of a gradient of replication fork polarity as in N-domains. However, in contrast to N-domains, U-domains are specific to one given cell line: U-domains are the somatic counterpart of N-domains. N-domains are thought to reflect the germline replication program since they are imprinted in the genomic sequence. Indeed, only the mutations occurring in the germline are transmitted and can accumulate to create the skew N-domains [106]. A recent study aiming at automatically detecting U-domains in seven human cell lines showed that U-domains cover roughly 50% of the genome [94, 101]. U-domains were objectively delineated using wavelet transform in the human genome for seven cell lines [101]. First, sharp peaks of the MRT profile, corresponding to initiation zones, were detected by finding regions of strong positive curvature (Eq. (II.4)). Subsequently, in between timing peaks, we retained only the domains where the MRT profile had a global negative curvature (*i.e.* was U-shaped). These genomic domains, U/N-domains, were shown to exhibit a striking gene organization and chromatin landscape [94, 104, 112, 114, 133, 136]. U/N-domains are units of the spatio-temporal replication program that share the same overall organization and are the eukaryotic equivalent of replicons. Therefore, they are preferential tools to study replication genome-wide.

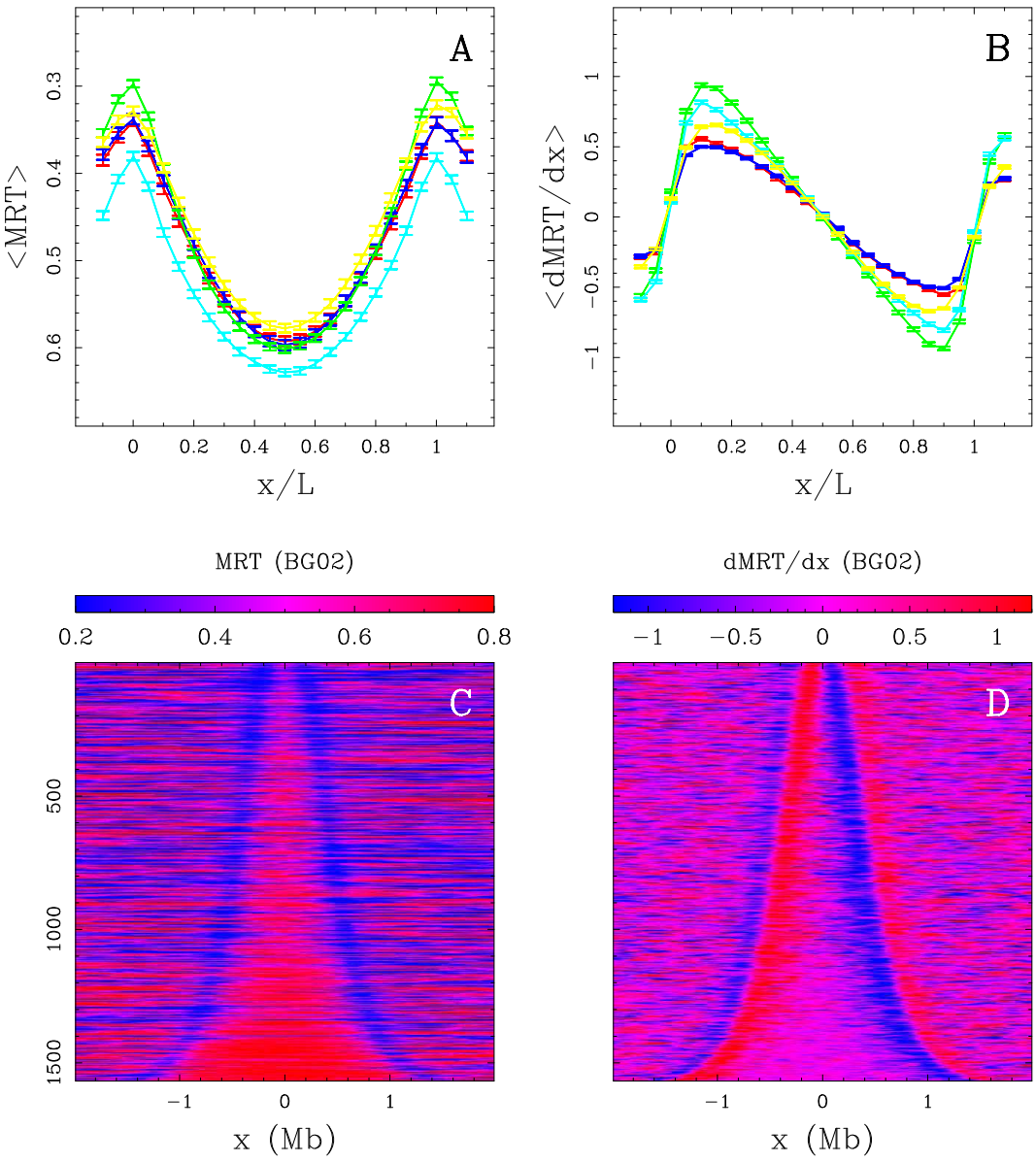

Figure II.7: Replication timing U-domains in different human cell lines. (A) Average MRT profiles (± SEM) inside detected replication U-domains. (B) Corresponding average $\frac{d\mathrm{MRT}}{dx}$ profiles (± SEM). In (A) and (B), each cell line is identified by a color: BG02 (green), K562 (red), GM06990 (blue), BJ R2 (magenta), and HeLa R2 (cyan). (C) The 2534 BG02 U-domains were centered and ordered vertically from the smallest (top) to the longest (bottom). The MRT profile of each domain is figured along a horizontal line using the MRT (BG02) color map. (D) Same as in (C) but for using the $\frac{d\mathrm{MRT}}{dx}$ (BG02) color map. Reproduced from [94].

II.2 Chromatin structure and replication

II.2.1 Chromatin is formed by successive folding layers

The DNA of eukaryotic cells is enclosed in the cell nucleus. Generally, eukaryotes have their genome organized in several separated chromosomes. Yet, even in separated chromosomes, the length of the longest DNA molecule in eukaryotic genome far exceeds the diameter of the cell nucleus which is on average 6 μm for mammalian cells. For instance the longest human chromosome[6] of 2.8×10^8 base pairs, which is almost 10 cm in length [137, 138]. The length of eukaryotic genomes implies two contradictory imperatives [4]. The genome must be condensed in such a way that it fits inside the nucleus while being highly organized so that every nuclear function (*e.g.* replication, mitosis, transcription) can take place efficiently.

The organized compaction must also be highly dynamic. Indeed, the compaction fold is 10000 during mitosis (chromosome must be tightly condensed to enable their proper distribution between the two daughter cells) and "only" 300 during the interphase[7]. Moreover, during the interphase, the compaction of the genome is heterogeneous. Transcriptionally or regulatory active regions are accessible and decondensed (euchromatin) while inactive regions and gene deserts are condensed (heterochromatin). The level of compaction during the interphase has to be adjustable to react to environmental cues.

This high degree of organization coupled to a tight compaction is obtained by the association of DNA with proteins. The complex formed by the DNA and the proteins attached to it is named **chromatin**. Actually, the weight of the proteins associated to DNA equals the weight of the DNA alone [138]. Chromatin is organized in successive layers of folding of increasing scale that are depicted in Fig. II.8. Each layer has its functional relevance and carries regulatory information [5, 6].

The first layer is the simple DNA helix. This is not really a chromatin layer since it does not contain proteins but it already holds epigenetic information. CpG dinucleotides can bear a methylation on the C nucleotide. This modification is present in inactive regions. On the second layer DNA wraps

[6]The longest human chromosome is chromosome 1. Human chromosomes are numbered from the longest to the shortest.

[7]The "compaction fold" is the ratio of the actual length end to end of the chromosome in the nucleus over the length of the chromosome DNA laid out as a perfect double helix.

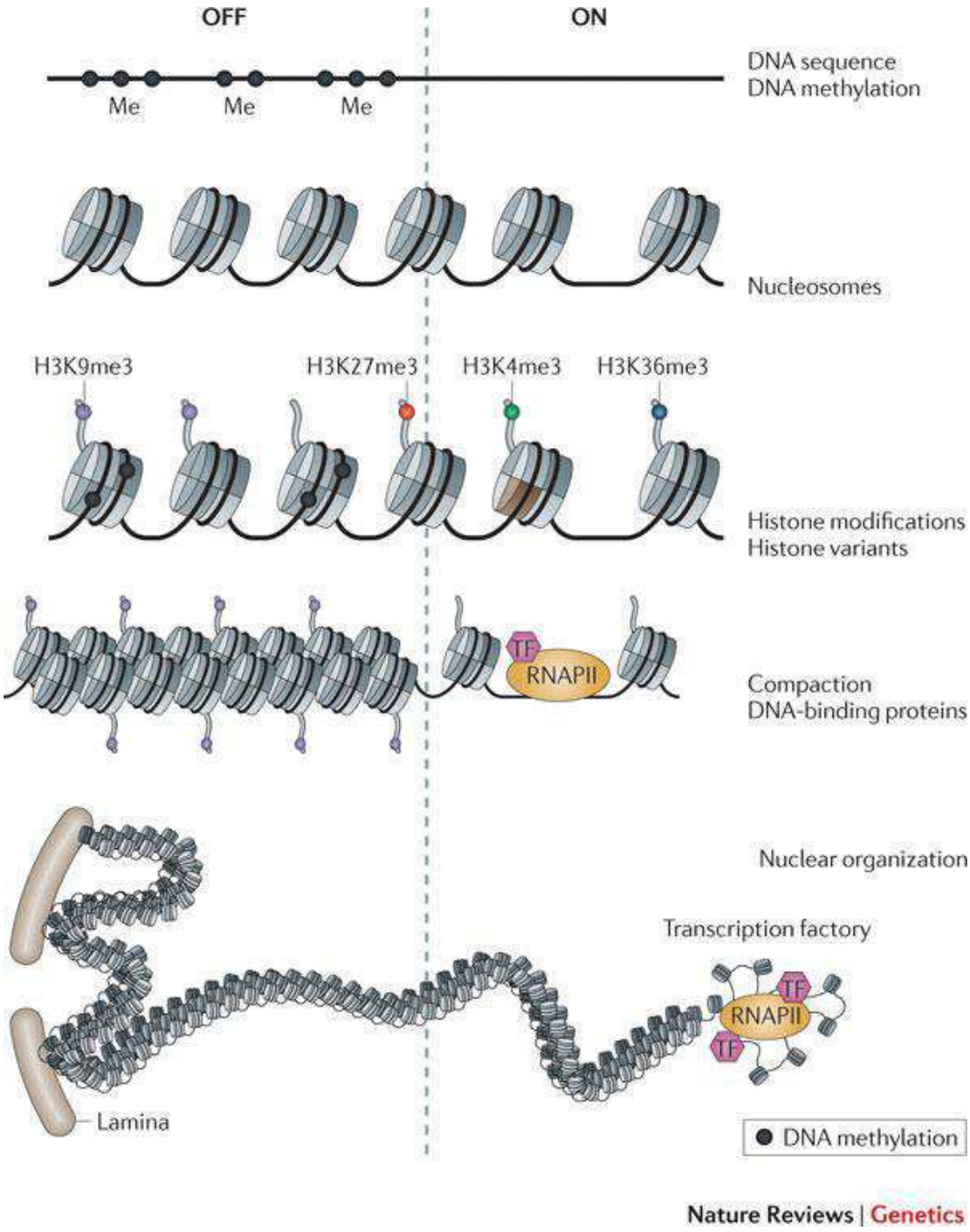

Figure II.8: Chromatin structure layers related to transcriptional activity. On the left part of the drawing, features associated with inactive regions (heterochromatin) are depicted; on the right, active regions (euchromatin) are described. See text for the detailed description of each layer of the chromatin. Sketch reproduced from [6]

.

around a bead-like structure formed by an octamer of proteins called histones. DNA turns about twice around each octamer (as pictured on the second row in Fig. II.8) [139–141]. The complex formed by the DNA and the eight histones is called nucleosome. Each canonical nucleosome contains 147 bp of DNA. The simple association formed between nucleosomes and DNA is a second layer of compaction called “bead-on-a-string". Nucleosomes can organize further in a third layer as a fiber of 30 nm whose exact structure is still debated. In active regions, chromatin has the least compacted form *i.e.* bead-on-string structure while inactive regions are condensed in a 30 nm fiber. Finally, the three-dimensional organization of DNA and nucleosomes is referred to as the fourth layer of chromatin. Chromatin has a specific spatial distribution in the nucleus: active regions are at the center and gene deserts are attached to the nucleus periphery with lamina fibers[8] [6].

Histones are the most prevalent proteins in chromatin. There are four types of histones: H3, H4, H2A, and H2B. H3 and H4 associate together to form a dimer; two dimers of H3-H4 associate to form a tetramer which is, in turn, surrounded by two dimers of H2A and H2B. Histones convey a lot of epigenetic information by two mechanisms. First, one of the canonical histones can be replaced by a histone variant[9]. This replacement slightly modifies the nucleosome structure and thus its function. For instance, transcriptionally active regions are enriched in the H3.3 and H2AZ variants instead of H3 and H2A, respectively. Second, histones are formed of a globular part that constitutes the nucleosome core and of a flexible "tail" that reaches outside toward the nuclear environment (rows 3 and 4 in Fig. II.8). These tails carry diverse modifications that have a functional meaning. There is a specific annotation to indicate modifications: H3K9ac means that a histone H3 carries a modification on its ninth amino acid and that the modification is an acetylation. Histone modifications can make the chromatin looser (acetylation modifications) or can serve as an anchor to regulatory proteins. For instance, H3K27me3 is used as a docking station by the Polycomb Repressive Complex PRC1 that silences developmental genes[10].

[8]The nuclear lamina is a dense fibrillar network inside the nucleus composed of intermediate filaments and membrane associated proteins.

[9]Proteins are formed by a chain of molecular units called amino-acids. Histone variants have the same amino acid sequence as the canonical histone with a couple of substitutions.

[10]To have a complete summary of the diverse histone modifications and their functions see the excellent review [55].

Other diverse proteins are associated to DNA. Some are needed to transcribe/ activate genes (transcription factors, enhancer proteins), some repair and replicate DNA, some repress genes by compacting the chromatin (HP1, polycomb) and others modify histones [5]. For instance the families of proteins that add and remove acetyl groups to histone tails are named HAT (Histone acetyl transferase) and HDAC (Histone deacetylase). There is also a class of proteins that move nucleosomes along DNA or eject them called chromatin remodelers. Finally, there is a class of proteins that organize the fourth layer of chromatin, namely its three dimensional folding (cohesin and CTCF). CTCF proteins form DNA loops that have various regulatory effects [142].

The successive layers of folding and the diverse proteins interact together to form the chromatin structure. In this thesis, we focus on the positioning of histone modifications and DNA binding proteins along the human genome. The influence of chromatin structure on the regulation of transcription has been widely assessed. However, chromatin structure serves other functions. We now describe how chromatin structure can influence DNA replication.

II.2.2 Chromatin structure and its influence on DNA replication

In this paragraph, we state the principles by which chromatin structure could be causally linked to replication. We will not make an exhaustive review on the subject. The interested reader should consult the excellent review [62].

Even though the DNA sequence may play a role in the positioning of origins [77,124], there is no consensus sequence for replication origins in metazoan. Origin positioning is cell line specific: even if there was a consensus sequence, an additional regulation mechanism would be needed [124]. Therefore, mechanisms that position and control the time of firing of origins must be epigenetic and linked to chromatin structure [61, 63–67]. Chromatin primary structure can influence replication by facilitating or preventing ORC deposition and origin firing. Indeed, ORC shows little or no sequence specificity in mammals, which indicates that its deposition must be epigenetically regulated. Experiments demonstrate that origin firing depends on chromatin environment [62]. Moreover, ORC positioning has been linked to nucleosome dynamics and can be efficiently predicted by a model taking only the enrichment in chromatin remodelers as a parameter [127]. A recent mapping of ORC in human revealed

that early origins are close to CpG-rich promoters displaying the H3K4me3 modification [128]. Another possible causal link not explored until today is the influence of chromatin compaction on the replication fork progression. The replication fork is probably slower in compacted region than in euchromatin regions.

Conversely, replication can influence chromatin structure by influencing its transmission. How histone marks subsist after successive replications has been a long standing question. Even though the mechanisms of inheritance of histone marks through replication remain to be clearly established [143], there are solid presumptions that histone modifications participate to epigenetic memory [144]. If histone marks were not transmitted through replication, then they would not be properly called epigenetic[11] marks [143]. Replication influences chromatin primary structure by reorganizing histones. To allow the progression of the replication fork, chromatin is temporarily disrupted, then the parental and newly synthesized histones are deposited on DNA [62, 65]. The latter action is controlled by histone chaperones that associate with the replication fork (Chromatin assembly factor 1 CAF-1 for H3 H4 dimer and nucleosome assembly protein 1 NAP-1 for H2A-H2B). A first mechanism of transmission could be that parental histones are redistributed between the two DNA strands."Old" histones would keep their modifications intact and serve as templates for newly synthesized histones [62]. An alternative mechanism to explain the epigenetic memory has been proposed in [144]. Histone modifications would be erased during the passage of the replication fork but histone modifying complexes could stay attached to the DNA thread. Behind the replication fork, these complexes would reestablish the histone marks. On a larger scale, it seems that replication may also act upon the three dimensional chromatin structure. Schematically, replication starts at the center of the nucleus and goes towards the periphery. During this progression, it can reorganize the 3D distribution of histone marks [145].

MRT data measure precisely the moment of the S-phase were the chromatin is disrupted and reassembled. These data seem more suited to study the causal effects of the replication program on chromatin structure. By contrast, the influence of chromatin on replication is more difficult to estimate. To precisely assess the effects of chromatin structure, we would need the intrinsic position-

[11] Epigenetics is the study of **heritable** changes that cannot be explained by changes in DNA sequence.

ing and firing times of origins which is impossible to measure directly because of passive replication[12]. Therefore, theoretical efforts are needed to delineate active and passive replication, which would enable to assess the genome wide effects of chromatin on replication initiation [131, 132, 146, 147]. Alternatively, *in vitro* experimental procedures could be proposed to assess the intrinsic firing time of replication origins.

Remark: Histone variants are deposited throughout the cell cycle and are called **replication independent** *which does not mean that they cannot impact replication. In the causality mechanisms discussed above, it would mean that replication does not impact them directly but, in reverse, they can impact replication by favoring or preventing origin deposition and firing. In Chapter V, we will see that H2AZ (replication independent) seems to be an important protagonist in the early firing of ESC specific replication origins.*

II.3 Data

This short section gives a few indications on how datasets used in the "results" chapters were produced.

II.3.1 CHiP-seq assay

To study the impact of chromatin on diverse nuclear functions, the fluctuations of its composition along the genome have to be known. To our knowledge, there are currently two different procedures to locate proteins on the DNA sequence: DamID [148] and CHiP-seq [149, 150]. The data used in this thesis have been produced by CHiP-seq, which stands for chromatin immunoprecipitation with massively parallel DNA sequencing.

Here, we briefly describe how protein positions are obtained by CHiP-seq (Fig. II.9):

* DNA and chromatin proteins are bond together by covalent links.
* Then DNA-protein complexes are sheared: naked DNA is digested thus DNA-protein complexes are separated from one another.

[12] A replication origin is said to be passively replicated if it is replicated by a fork coming from a neighboring origin.

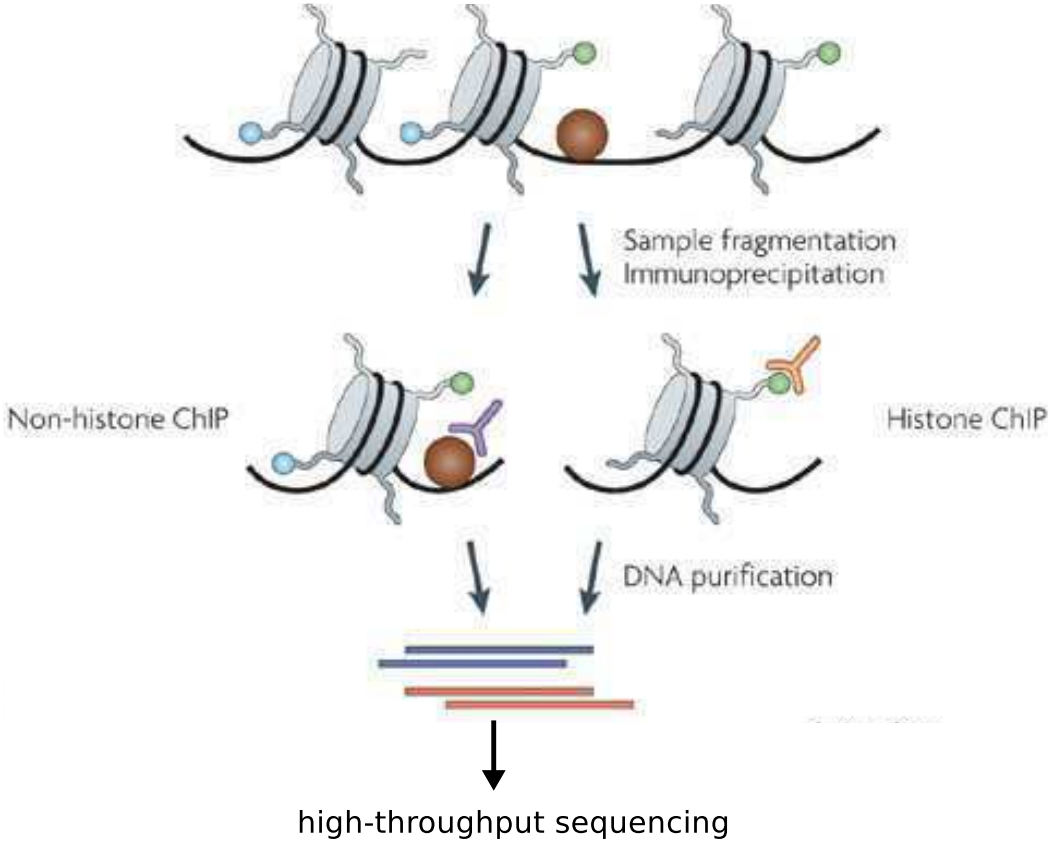

Figure II.9: Summary of the CHiP-seq protocol. Reproduced from [149].

* The complexes of interest (one particular protein, *e.g.* one histone bearing a modification or a DNA binding protein) are selected by an antibody. The antibody attaches to its antigen (DNA-protein complex) and precipitates.

* From the precipitate, the DNA is purified.

* DNA is sequenced by a high-throughput sequencer. Reads are then mapped on the genome.

Once reads are mapped on the genome, statistical pipelines detect positions of significant enrichment compared to the background noise. The final output of the procedure is a set of genomic intervals where the protein of interest binds.

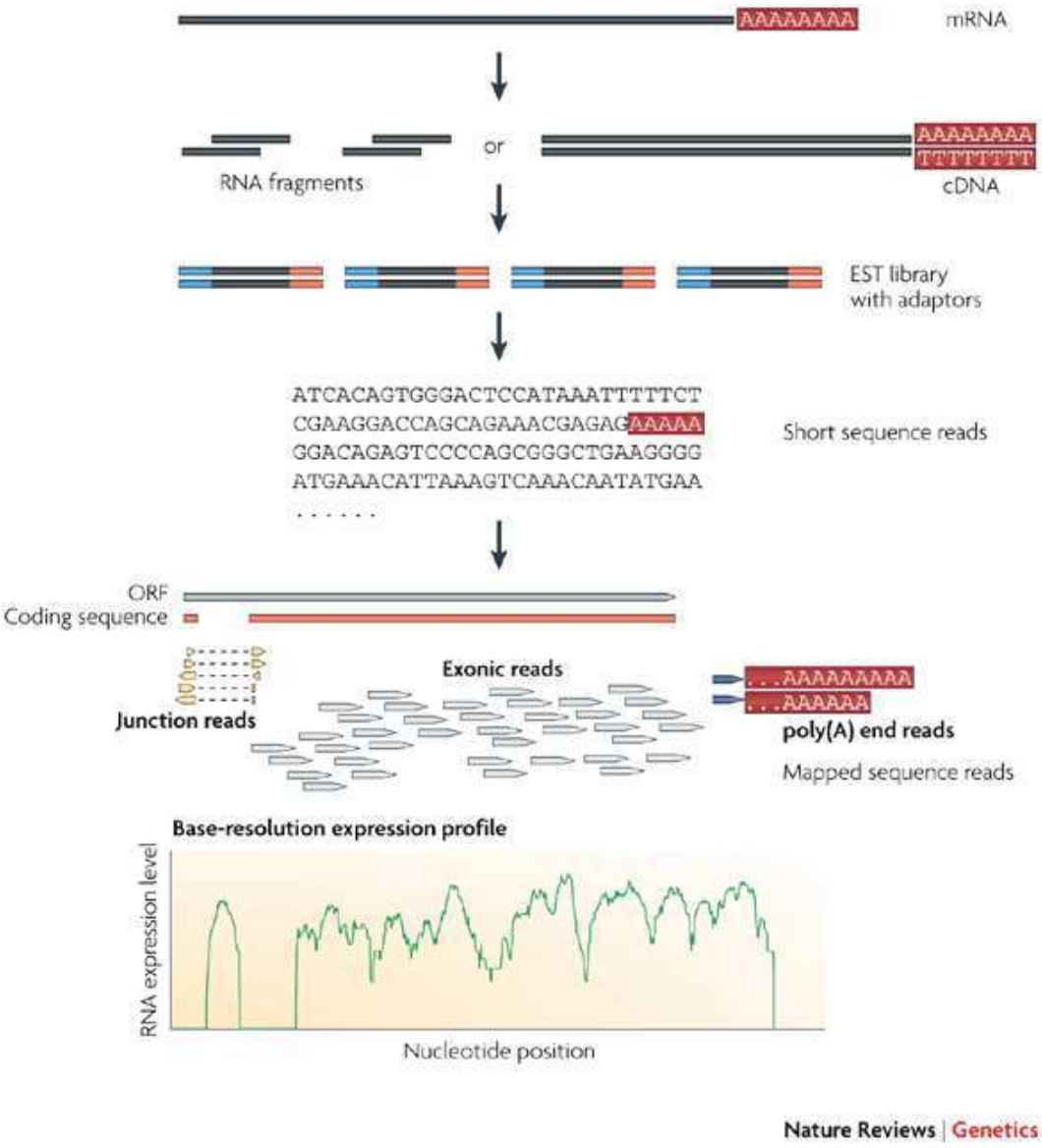

Figure II.10: RNA-seq protocol summary. Reproduced from [151].

II.3.2 Assessment of gene expression level by RNA-seq

RNA-seq is a transcriptomic [13] tool taking advantage of the high-throughput sequencing technology [10, 151]. It consists of selecting the RNA population of

[13] Transcriptomics is the study of the complete set of transcripts (*i.e.* RNA molecules) in a cell.

interest (*e.g.* small RNAs, messenger RNAs, non coding RNAs) and sequencing it. In our dataset, RNA from protein coding genes were selected. They were detected using their poly AAA tails.

Briefly, RNAs were first converted into a library of cDNA. In Fig. II.10, sequencing adaptors (blue) were subsequently added to each cDNA fragment and a short sequence was obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads were aligned with the genome.

One of the main advantages of RNA-seq compared to microarray technology[14] is that it does not rely on a gene annotation [151]. It can be used to complete annotation (discovering new genes, improving positioning of exon-intron boundaries). The RPKM unit defined in the "results" chapters (see *e.g.* Sect. III.4.8, page 88) is equivalent to counting the number of times the RNA polymerase transcribes a gene. This measure is far more natural and does not present the same limits (saturation, noise...) as fluorescence intensity used in microarrays. It also presents the advantage of being readily comparable between experiments which was difficult to do for microarrays without complex normalization methods [151].

II.4 Statistical methodology

This section is dedicated to the description of statistical techniques used in the "results" chapters. The analyses presented in this thesis use successively three statistical tools: the Spearman correlation, the Principal Component Analysis (PCA) and an optimized K-means clustering (Clara). The principles of these techniques are presented here in a way understandable without statistical knowledge. However we would also like to provide a theoretical explanation to the interested reader. To solve this apparent paradox, each subsection deals with one technique and is organized as follows:

* Presentation of the purpose of the technique.
* Definition/algorithm of the technique. The non-statistician may skip this paragraph.

[14] A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Microarrays are used to measure the expression levels of large numbers of genes simultaneously by specifying a probe for each gene of interest.

* An example on a simple dataset to illustrate the power and interest of the tool. For the clustering and PCA subsections, the example treated is treated as the dataset of interest in the "results" chapters.

II.4.1 Pearson correlation and Spearman correlation

Purpose: Correlation coefficients are basic tools widely used. On a population of observations/individuals/items (*e.g.* genes, 100 kb windows, ORC binding sites), one can measure diverse variables that can be continuous (*e.g.* RPKM expression level, length, $CpG_{o/e}$) or categorical (CpG rich/CpG poor). Correlation is computed between two variables and assesses to what extent the variables vary simultaneously. If the variables increase at the same time they are said to correlate. Inversely, if one variable increases while the other decreases, they anticorrelate. However, the classical Pearson correlation coefficient measures if a variable linearly increases with another variable. To our judgment, there is no particular interest to look for a linear relations between variables. The Spearman correlation coefficient measures how the rank of a variable varies with the rank of another variable[15]. Consequently, the Spearman correlation captures all kind of monotonical relation between two variables.

Definition: We remind that the Pearson correlation coefficient, between two random variables X and Y, is given by [152]:

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}. \tag{II.5}$$

For n observations (x_i, y_i) of the couple of random variables (X, Y), the empirical correlation is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{II.6}$$

where $\bar{x}$ and $\bar{y}$ are the empirical means.

The Spearman correlation [152] is simply the Pearson correlation between the ranks of the variables rather than the variable values themselves. Let r_i

[15]The rank of an observation is its position, according to its value, in the ordered list of all observations. Therefore the smallest value has rank 1 and the largest value has rank N.

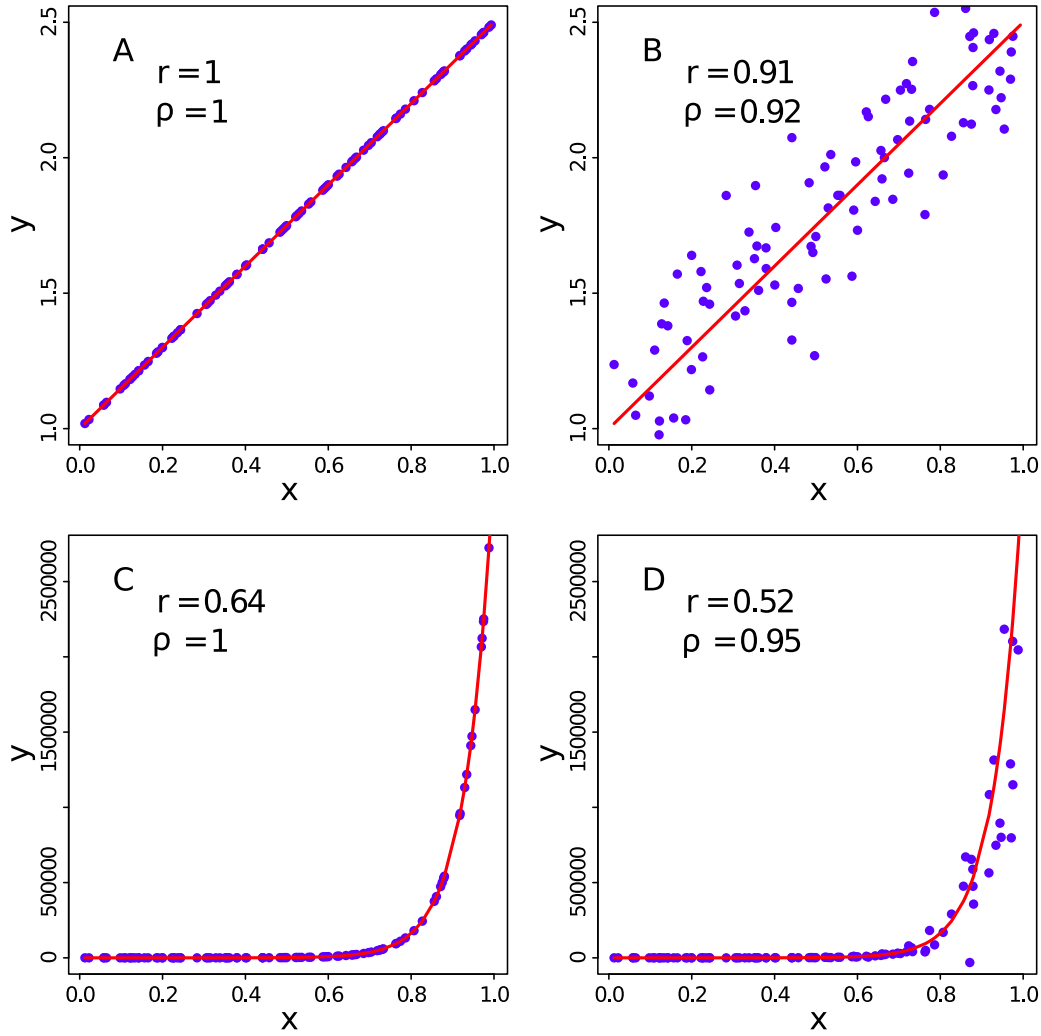

Figure II.11: Difference between the Spearman and the Pearson correlation coefficients. Each panel presents a particular relation between two variables x and y. In the top-left corner of each panel the Pearson correlation coefficient (r) (Eq. (II.6)) and the Spearman correlation coefficient (ρ)(Eq. (II.7)) are given. (A) Linear deterministic relation $y = 1.5 \times x + 1$. (B) Same relation as A with noise. (C) Exponential relation between x and y : $y = exp(15x)+1$. (D) Same as (C) with noise. The red line represent the deterministic relation between y and x. Blue dots represent the actual data set on which the correlaton coefficients were calculated.

be the rank of x_i and s_i the rank of y_i:

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(r_i - \bar{r}) \times (s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2} \times \sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}, \tag{II.7}$$

where $\bar{r}$ and $\bar{s}$ are the empirical means.

Example: To illustrate the difference between the Spearman and Pearson correlation coefficients, we consider as examples a perfect linear relation and an exponential relation between two variables (Fig. II.11). In the first case, these correlation coefficients are equal (Fig II.11A). The two coefficients capture efficiently linear relations between two covariates. Moreover, they are robust to a reasonable amount of noise (Fig. II.11B). The second example is constructed with the same x data points. The y data points are an exponential function of x. In the deterministic case, the Spearman correlation value is still 1 whereas the Pearson correlation coefficient decreases to ~ 0.6 (Fig. II.11C). Since the relation is deterministic and monotonical, it seems desirable that the correlation coefficient value be maximal (*i.e.* 1). The Spearman coefficient clearly indicates that the value of x gives the value of y. When noise is added, the Spearman correlation coefficient is still high ~ 0.9 while the Pearson coefficient is only 0.5 (Fig. II.11D) which seems weak for such a clear relation between variables.

Remark: the CHiP-seq read density spans several order of magnitude as the y data points in the preceding exponential example. Actually, the Pearson correlation is thus particularly ill-suited to the kind of dataset analyzed in the "results" chapters. For these practical and theoretical reasons, we choose to use the Spearman correlation coefficient over the Pearson coefficient.

II.4.2 Principal Component Analysis

Purpose: Principal Component Analysis (PCA) is a dimensionality reduction technique [152]. Its purpose is to estimate the "true" dimensionality of a data set. Indeed, a dataset that appears to have a high number of dimensions[16] because a lot of correlated variables were taken into account (like in the introductory example with DamID 46 profiles) can have a "real" dimension of only two or three. Imagine data points aligned in a 3D space. The line where the data points are confined is one-dimensional and the position of the line could be properly described by only one vector. The position of the data points are described equally well by their coordinates along this line and by their three coordinates in the original space. The projection has interesting

[16]The number of dimensions is the number of variables measured on the individuals.

properties since it contains in only one coordinate the information contained in three coordinates in the original dataset. *PCA is a technique that finds proper vectors to project data points on and to estimate how well the projection represents the original data.* In practice, data points are never exactly confined to a smaller dimensional space. Therefore the projection does not exactly reflect the original data set. For the PCA to be meaningful, the information lost by projection should be noise or irrelevant.

Because PCA makes dataset smaller, it can be used for data compaction. Alternatively, PCA can be used to generate more manageable datasets than the original one. For instance, PCA projections are useful to visualize datasets of high dimensionality.

Definition: PCA defines the projection vectors by diagonalizing the covariance matrix of a mutivariate dataset. If we measure D variables on N individuals, the dataset can be represented by a $(N \times D)$ matrix. Let X be this $(N \times D)$ matrix and $X_{i\alpha}$ the variable α for the individual i. Thus, a row of the matrix X is noted X_i and encompasses all variables for the individual i.

The mean vector (vector of the mean value of the D variables) is computed as :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{II.8}$$

We define a centered row as $XC_i = X_i - \bar{X}.$ and XC the matrix formed by all centered rows.

The sample estimate of the covariance matrix is

$$\hat{\Sigma}_{XX} = \frac{1}{n} XC.XC^t, \tag{II.9}$$

where XC^t is the transpose of XC. In other words, $\hat{\Sigma}_{\alpha\beta} = \frac{1}{n}\sum_{i=1}^{n} XC_{i\alpha}XC_{i\beta}$. It is clear that the covariance matrix is a positive symmetrical matrix. Linear algebra theorems state that such positive symmetric matrices are diagonalizable in an orthornomal basis and that all the eigenvalues are positive.

PCA consists in finding the eigenvectors of the covariance matrix. Projections on these eigenvectors are linear combinations of the original random variables. The new random variables defined by the eigenvectors are uncorrelated to each other and their variances are given by the corresponding eigenvalues. The decreasing eigenvalues of $\hat{\Sigma}_{XX}$ are denoted by $\hat{\lambda_1} \geq \hat{\lambda_2} \geq \cdots \geq \hat{\lambda_D} \geq 0$,

and v_k is the eigenvector associated to the k^{th} largest eigenvalue. Only a few vectors with the largest eigenvalues are retained and the original data are projected on this basis.

Since projections on each eigenvectors are uncorrelated, the variance of several projections is additive. Therefore, the variance of the projection on the space formed by (v_1, v_2, v_3) is $\lambda_1+\lambda_2+\lambda_3$. Statisticians think about information as variance. Suppose, for example, that in a multivariate dataset, one variable has the same value for all individuals. This particular variable does not truly contain information. Indeed, we can acknowledge that this feature is constant and remove it from the original dataset. Consistently with its information content, the variable variance is zero. By contrast, a random variable with a high variance is said to contain a lot of information because it differentiates the individuals from one another. PCA creates a set of uncorrelated variables, classifies them according to their variances and keeps only variables with a non-negligible variance ($i.e$ those that contain a lot of information).

If the j first eigenvectors are retained then the proportion of variance accounted for is

$$\text{Information} = \frac{\sum_{k=1}^{j} \hat{\lambda_k}}{\sum_{k=1}^{D} \hat{\lambda_k}}. \tag{II.10}$$

In practice, the proportion of explained variance is used to estimate the number of eigenvectors to retain. One can set a threshold and retrieve the number of eigenvectors necessary to explain 90% of the total variance. Alternatively, one can search for an interesting trade off between the number of eigenvectors and the variance explained.

Example: We illustrate the usefulness of PCA on the artificial example shown in Fig. II.12. The dataset is distributed in 5 different 3D multivariate normal distributions. To illustrate how PCA efficiently detects informational projection subspaces, we voluntarily constrain the mean of the multivariate distributions to the same plane and one variable contains only noise (z axis). In the original 3D space, the scatter plot of data points does not reveal at all the existence of the five multivariate modes (Fig. II.12A). The eigenvalues obtained by diagonalizing the correlation matrix are 18.25, 6.54, 0.32. Therefore, the first principal component accounts for 72.7% of the total variance, the second for 26.1% and the last for 1.3%. The information on the last component is negligible. Hence, only the first and second principal components are retained.

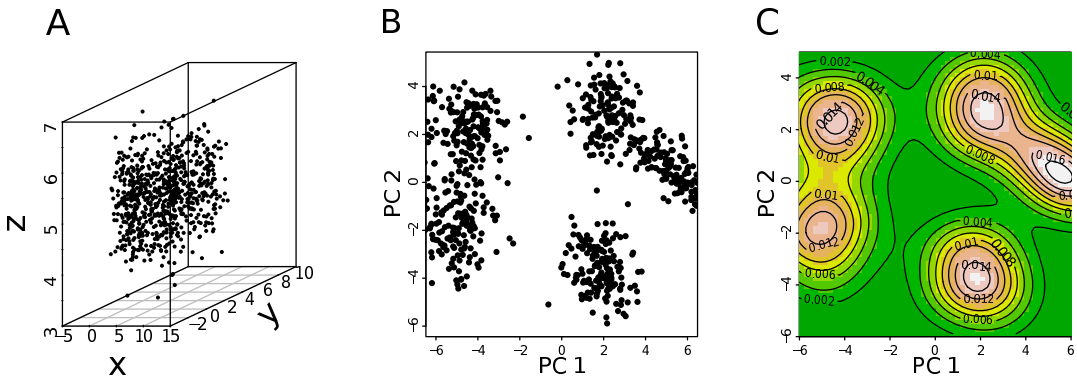

Figure II.12: Finding meaningful projections with PCA. (A) Scatterplot of the original dataset in 3D. (B) Projection of data points onto the plane formed by the first principal component (PC1) and the second principal component (PC2). (C) Same projection as in (B) represented as a density. The densities were computed by a kernel density estimator. The density values are indicated by a color (white: high density, yellow: moderate density, green: low density) and a contour plot.

The structure of the dataset appears clearer on the projection on the plane defined by the first (PC1) and second (PC2) principal components (Fig. II.12B). Three clouds are distinctly separable. Two of the three clouds may be divided in two smaller clouds. To decide on the number of “clouds", a better representation is needed. Indeed, scatter plots can be misleading because dots overlap. Therefore, scatterplots saturate and high density regions are not distinguished from moderate density regions. To visualize the law underlying the data point spatial distribution, one can use a non parametric density estimator like kernel estimators. The probability density function is estimated from the data points and represented in Fig. II.12C. The five original modes clearly appear even though the two in the top-right corner are a little bit overlapping. The underlying structure of data points is much more apparent after applying PCA and a good visualisation device (Fig. II.12C) than in the original 3D space (Fig. II.12A).

Note that for such a simple example, a spinning 3D scatterplot is probably enough to understand that the data contain 5 modes. However when the original dataset has a high dimensionality ($\gtrsim 10$) simple visual inspection is obviously impossible. In the “results” chapters, the original datasets, that contain more than ten variables, will be reduced, thanks to PCA, to a 3D projection. This will enable visual inspection which, from our standpoint, is a very good way to detect structures in a dataset.

II.4.3 Clustering

Purpose: A clustering method refers to an algorithm that objectively find homogeneous groups in a dataset. Each cluster (*i.e.* group of individuals) should contain observations that are similar to each other. Another desirable property is that observations in different clusters should be significantly different. There are a lot of methods to cluster objects and their outputs are not identical and can even be very different. Indeed, the clustering algorithm depends heavily on the definition of similarity between individuals. The clustering method must be chosen according to the data of interest. An issue with clustering is that it is difficult to assess the quality of a clustering because it is an unsupervised method (see Chapter VI). There are a lot of *ad hoc* criteria to estimate the quality of a clustering on a given dataset. Yet, different criteria can lead to the choice of different "optimal" clustering methods.

Once the initial dataset is cleared of negligible dimensions, we can visualize our data and see if there are detectable high density zones in it. In this thesis we favor an approach based on visualisation (enabled by PCA) and the analysis of the output over *ad hoc* criteria. We select the right clustering method by verifying the concordance between clusters and visible high density regions in the data set. Finally, the clustering quality is confirmed *a posteriori* by the relevance of the obtained results from a biology standpoint.

Algorithm: We used the Clara algorithm [153] which is an optimization of k-means for large datasets. According to k-means algorithm, let K be the number of clusters in the dataset. The centroid of a cluster is the mean position of individuals belonging to this cluster. If X_i is the vector of coordinates for the i^{th} individual, then the cluster centroid $\bar{X_k}$ of the k^{th} cluster is defined as:

$$\bar{X_k} = \frac{1}{n_k} \sum_{j \in k} X_j, \tag{II.11}$$

where n_k is the number of individuals in the k^{th} cluster.

1. The input is the dataset $X = \{X_i, i = 1, 2, \ldots, n\}$ and the number of clusters K.

2. Initialize the procedure by doing one of the following:

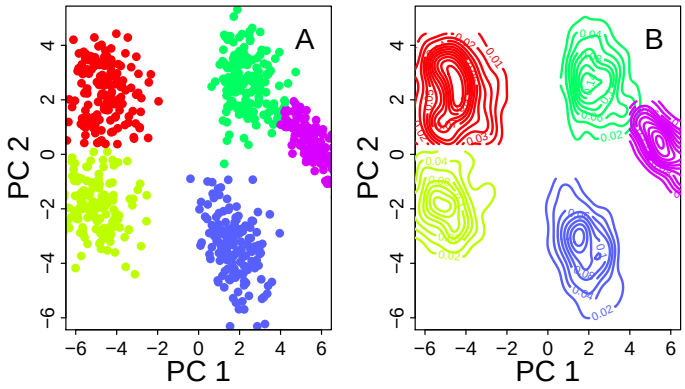

Figure II.13: Clare clustering method delineates high density regions. (A) Clustering of data points in the plane formed by the first (PC1) and the second (PC2) principal components. Each color corresponds to a cluster. (B) Contour plots of the densities of the five clusters computed by a kernel density estimator.

   - Assign each observation to a cluster randomly. For each cluster compute its centroid.
   - Specify K cluster centroids, $\bar{X}_k$, $k = 1, 2, \ldots, K$

3. Compute the squared-euclidean distance of each item to the K cluster centroids and assign items to the closest centroid.

4. Compute the new positions of centroids according to the new assignation of items.

5. Repeat step 3 and 4 until the assignations do not change.

The complexity of k-means is $O(N^2)$ which can be an issue for large data set. The trick used by Clara (Clustering for Large Application) [153] to optimize k-means is the following:

- Draw several samples of a smaller size than the original dataset.
- Estimate the centroid positions by running k-means on the small samples.
- Assign individuals according to the estimated centroid positions.

Since the size of the sample is constant, the complexity does not increase as a quadratic function.

Example: Fig. II.13 shows how Clara efficiently segregates the five laws underlying in the 3D distribution of data points in the example discussed in Fig. II.12. Clara was applied on the projection of the dataset on the plane formed by the first (PC1) and second (PC2) components (Fig. II.12B). The output of the Clara clustering is displayed in Fig II.13A. The kernel estimation (Fig. II.13B) of each cluster distribution demonstrates that to each cluster corresponds a mode of the original data point distribution (Fig. II.12C).

In a nutshell, the workflow we repeatedly apply in the "results" chapters can be summarized as follows. PCA enables a visual inspection of the data set. This inspection assesses the number of particularities we would like to detect. Then, the clustering algorithm automatically delineate groups that correspond to these particularities. By visual inspection, we make sure that the algorithm correctly detect the particularities.

Chapter III

Human Genome Replication Proceeds Through Four Chromatin States

In this Chapter, we perform the integrative analysis of the genome-wide distributions of thirteen epigenetic marks in the human cell line K562, at the 100kb resolution of corresponding mean replication timing (MRT) data. We thereby identify four major groups of chromatin marks. These states have different MRT, namely from early to late replicating, replication proceeds through a transcriptionally active euchromatin state (C1), a repressive type of chromatin (C2) associated with polycomb complexes, a silent state (C3) not enriched in any available marks, and a gene poor HP1-associated heterochromatin state (C4). When mapping these chromatin states inside the megabase-sized U-domains (U-shaped MRT profile) covering about 50% of the human genome (Sect. II.1.4), we reveal that the associated replication fork polarity gradient corresponds to a directional path across the four chromatin states, from C1 at U-domains borders followed by C2, C3 and C4 at centers. Analysis of the other genome half is consistent with early and late replication loci occurring in separate compartments, the former corresponds to gene-rich, high-GC domains of intermingled chromatin states C1 and C2, whereas the latter corresponds to gene-poor, low-GC domains of alternating chromatin states C3 and C4 or long C4 domains. This segmentation sheds a new light on the epigenetic regulation of the spatio-temporal replication program in human and provides a framework for further studies in different cell types, in both health and disease. Results

reported in this chapter are published in [154].

III.1 Introduction

Understanding the role of chromatin structure and dynamics in the regulation of the nuclear functions including transcription and replication, is a major challenge of current research in genomics and epigenomics [4, 6, 54–57, 155]. Since the initial sequencing of complete genomes and more than a decade ago of the human genome [1], the development of new techniques, in particular chromatin immunoprecipitation (ChIP) followed by massive parallel sequencing (ChIP-seq) [150], has enabled genome-wide analysis of many epigenetic modifications such as histone modifications, histone variant incorporation as well as of various DNA-binding proteins [6]. These techniques have been extensively applied to various eukaryotic genomes, from budding yeast [156], to plants [157, 158], worm [83], fly [159, 160], mouse [6, 9, 15] and human [6, 9, 15, 16], and have led to significant progress in our understanding of the chromatin landscape and of its impact on gene regulation, replication origin specification and cell differentiation. Statistical analyses of these multivariate data sets have shown that this huge combinatorial complexity can be reduced to a surprisingly small number of predominant chromatin states with shared features namely four in *Arabidopsis thaliana* [51], five in *Caenorhabditis elegans* [52] and four [53] or five [50] in *Drosophila*. To our knowledge, no such a drastic dimensional reduction has been reported in mammalian organisms so far. The application of a multivariate Hidden Markov Model (HMM) [48] as well as the implementation of adapted pattern-finding algorithm [161], have confirmed that distinct epigenetic modifications often exist in well-defined combinations corresponding to different genomic elements like promoters, enhancers, exons, repeated sequences and/or to distinct modes of regulation of gene expression such as actualy transcribed, silenced and poised [48,161–163]. Some recent study [164] of chromatin mark maps across nine different human cell types has ultimately identified fifteen main chromatin types which is a reljatively limited number of epigenetic states but probably not the optimal complexity reduction one may achieve in human and more generally in mammalian genomes. The analysis of a wide set of chromatin regulators that add, remove or bind histone modifications reported in [165], is a very encouraging step in this direction since six major groups or modules of chromatin regulators were shown to encompass the

combinatorial complexity and to be associated with distinct genomic features and chromatin environments. Here, we perform principal component analysis (PCA) [152] and classical clustering [153] on thirteen epigenetic mark maps in the K562 immature myeloid human cell line at the resolution 100 kb of corresponding available MRT data, with the perspective of identifying the major types of chromatin states in relation with replication timing during S-phase.

III.2 Results/Discussion

III.2.1 Combinatorial analysis of chromatin marks

We investigated relationships between the genome-wide distributions of eight histone modifications, one histone variant and four DNA binding proteins in the immature myeloid human cell line K562 (Sect. III.4.2) at the 100kb resolution of corresponding MRT data [14, 94]. As a first step, we computed the Spearman correlation coefficient (Sect. III.4.4) of each mark with each other. We next represented the resulting matrix as a heat map after having reorganized rows and columns with a hierarchical clustering based on the Spearman correlation distance (Eq. (III.1), Fig. III.1). This preliminary analysis was very promising as regards to the possibility of reducing combinatorial complexity. All the epigenetic marks that are known to be involved in transcription positive regulation, namely H4K20me1, H3K9me1, H3K4me3, H3K27ac, RNAPII, CBX3, H2AZ, H3K79me2, H3K36me3, together with the transcription factors CTCF and Sin3A, form a block in the correlation matrix, meaning that they are all correlated with each other. The maximum correlation is actually obtained between the two active promoter marks H3K4me3 and H3K27ac. As suggested in Refs [164, 166], all these active marks are likely to occupy similar regions in the genome. In fact, two lines are clearly apart on the hierarchical clustering dendrogram (Fig. III.1). They correspond to the repressive chromatin marks H3K27me3 and H3K9me3 that are respectively associated with the so-called facultative and constituve heterochromatins [145,166]. These two marks are recognized by the chromodomains of polycomb (Pc) proteins and heterochromatin protein 1 (HP1), respectively, components of distinct gene silencing mechanisms which likely explains that they are strongly anticorrelated with each other. While H3K9me3 behaves quite independently with respect to most of the active chromatin marks, H3K27me3 correlates to some of them

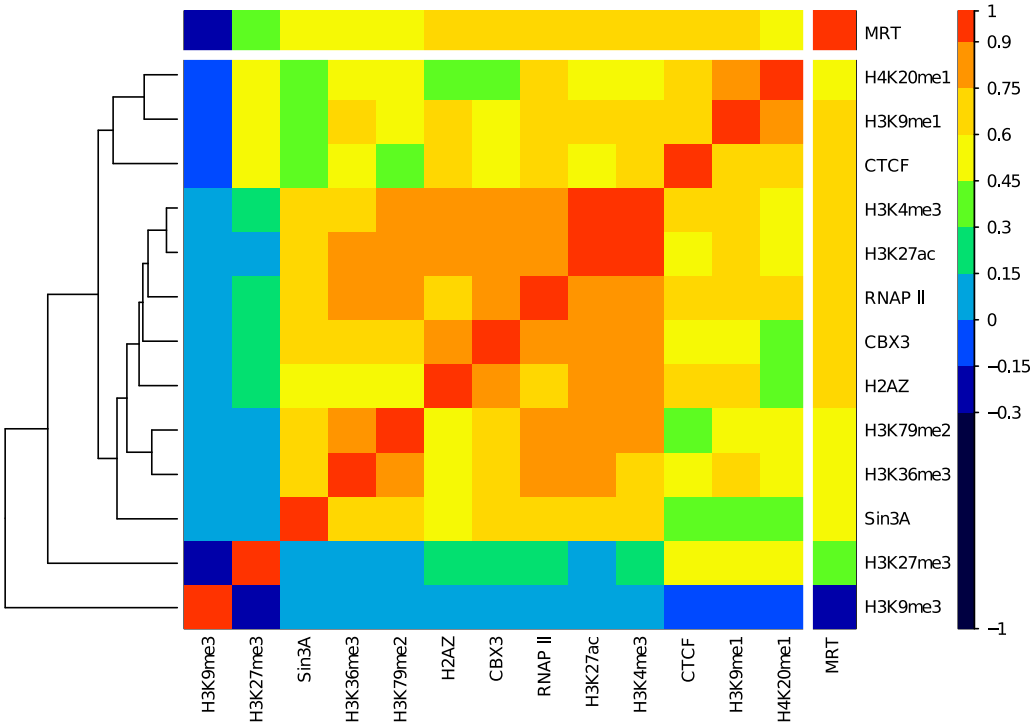

Figure III.1: Spearman correlation matrix between epigenetics marks and mean replication timing (MRT). For each pair of variables we computed the Spearman correlation over all 100 kb non-overlapping windows with a valid score. Spearman correlation value is color coded using the color map shown on the right. A white line separates the MRT from epigenetics marks. Correlations with MRT (from late to early) are placed at the top and the right of the matrix. Lines for the thirteen epigenetic marks were reorganized by a hierarchical clustering using Spearman correlation distances (Eq. (III.1)) as illustrated by the dendrogram on the left of the graph. This ordering implies that highly correlated epigenetic marks are close to each other.

and especially to H4K20me1, H3K9me1 and CTCF. When further investigating the correlations between the thirteen considered chromatin marks and the MRT (Fig. III.1), we found, consistently with previous works [12,14,34,36,89], a strong correlation for the transcriptionally active marks with early replication. Some moderate correlation was obtained for the Pc associated repressive marks H3K27me3 which contrasts with the significant anticorrelation observed for the constitutive heterochromatin mark H3K9me3 with late replication.

In a second step, to objectively identify the prevalent combinatorial patterns of the thirteen chromatin marks, we performed a PCA [167] to reduce the dimensionality of the data (Sect. III.4.5). We then concentrated on the

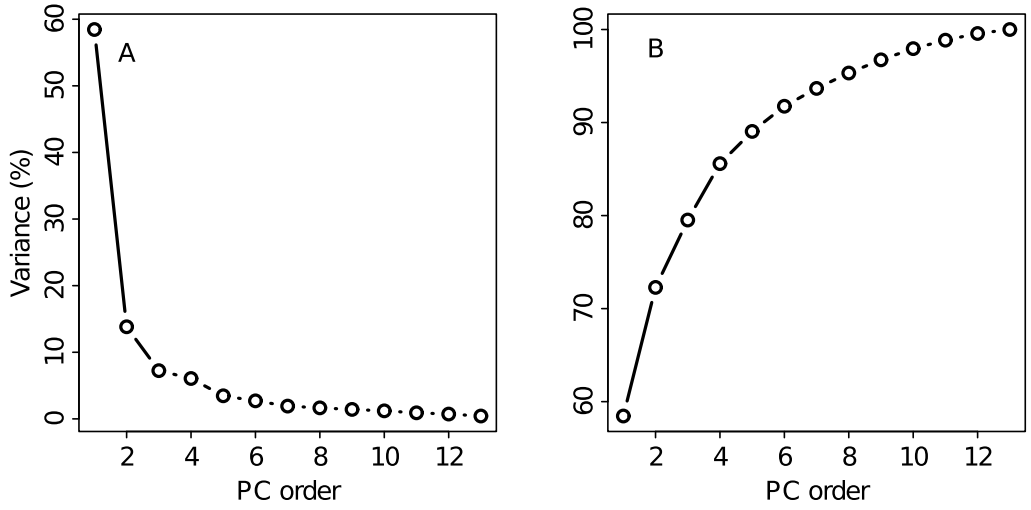

Figure III.2: PCA analysis. (A) Percentage of variance accounted by the first thirteen principal components ordered according to their corresponding variance (eigenvalues). (B) Cumulative variance.

first three principal components, which together account for 76% of the total data set variance (Fig. III.2). By projecting the 100kb genomic loci on the (PC1, PC2) plane (Fig. III.3A) and the (PC3, PC2) plane (Fig. III.3B), we noticed that four areas contain most of the population. On the (PC1, PC2) plane, a large area of medium density comes out from a plane of much higher density. As viewed on the (PC3, PC2) plane, in this very dense plane, loci mainly lie along two straight lines with a very high density of loci concentrated at the intersection of these lines. This led us to use the Clara clustering algorithm [153], which is very similar to k-means, with the number of clusters fixed to four (Sect. III.4.6). When labeling each of the four main chromatin states with a color, we obtained four domains in the 3D scatter plot (Fig. III.4A) that have common boundaries as evidenced on the three orthogonal projections on the planes (PC1, PC2), (PC1, PC3) and (PC3, PC2) (Fig. III.4A). To improve the quality of our clustering procedure, we filtered out poorly clustered data points that are closer to another cluster than to the one they belong to (black dots in Fig. III.4), where the distance between a data point and a cluster is defined as the mean of the distances of this point to all the points in the cluster. Removing those points is exactly equivalent as removing points with a negative silhouette [168] (Sect. III.4.6).

To determine the number of clusters, we used two statistical criteria (Sect.

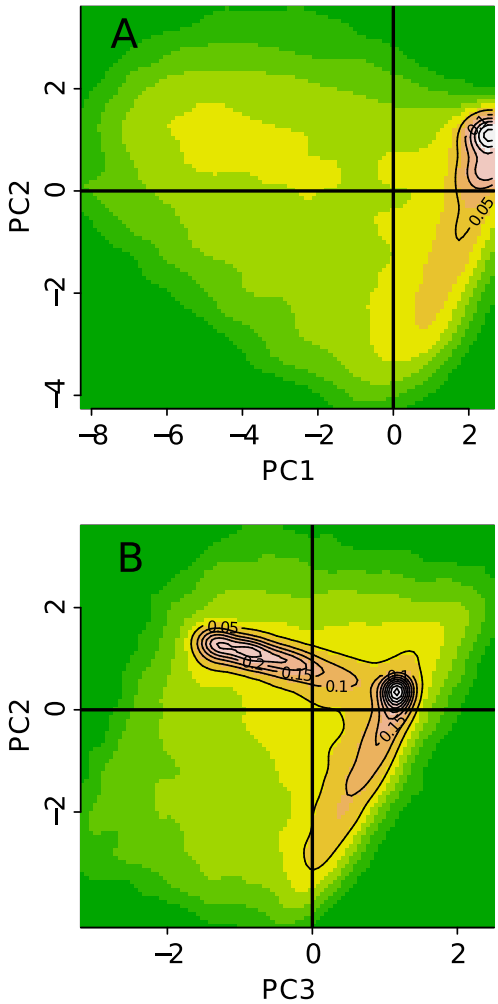

Figure III.3: Principal Component Analysis (PCA). Two-dimensional (2D) projections of the data on (A) the plane defined by the first (PC1) and second (PC2) principal components, and (B) the plane defined by the second (PC2) and the third (PC3)principal components. The densities were computed by a kernel density estimation. The density values are indicated by a color (white: high density, yellow: moderate density, green: low density) and a contour plot.

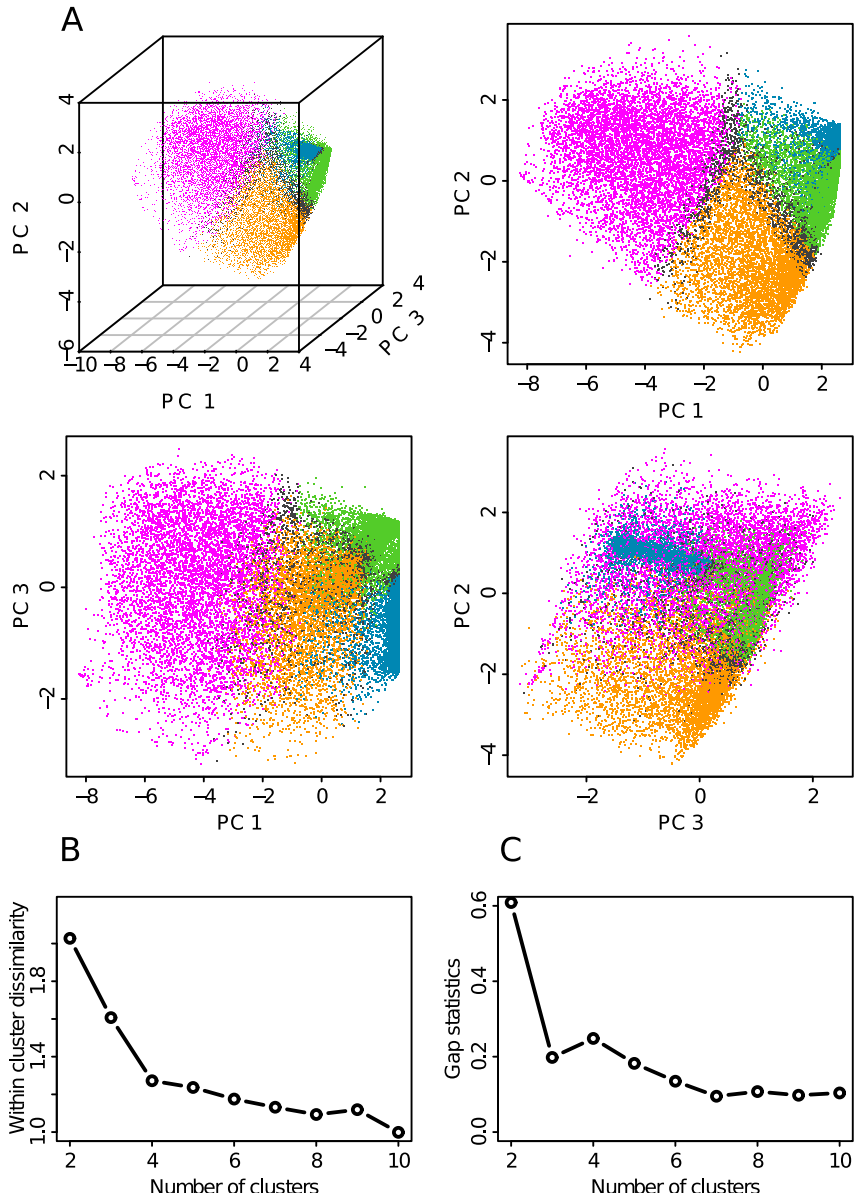

Figure III.4: Defining the four prevalent chromatin states. (A) Scatterplot of the data points onto the first three principal components. Color dots indicate the four chromatin states as found by our clustering procedure (pink: transcriptionally active chromatin, orange: chromatin repressed by polycomb, green: silent unmarked chromatin, blue: HP1 heterochromatin). Points in dark grey are not classified in any chromatin state (Sect. III.4.6). (B) Within-cluster sum of squares (Eq. (III.2)) with respect to the number of clusters. (C) Gap statistics (Eq. (III.4)) with respect to the number of clusters.

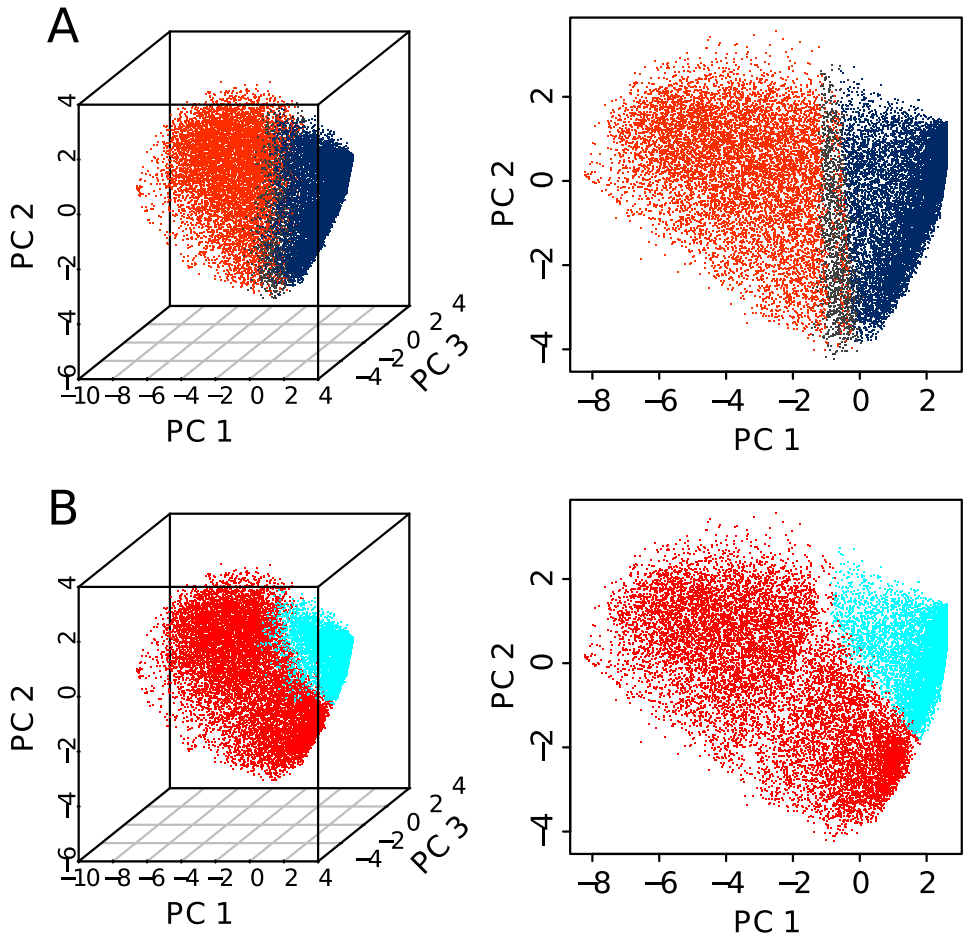

Figure III.5: Dichotomic analysis with two chromatin states. (A) Results of our clustering procedure when using two clusters (the number of clusters is the only parameter of the procedure). We found a segmentation between transcriptionally active chromatin (red) and silent chromatin (blue). (B) Same representation for chromatin state blocks (1+2) (light red) and (3+4) (light blue) as defined in Fig. III.19D.

III.4.6). Four is the optimal choice according to the within-cluster sum of squares that clearly displays an elbow (abrupt slowing down of the decay) at the cluster number equal to four (Fig. III.4B). The gap statistic [169] indicates that two or four clusters are good solutions (Fig. III.4C). Our choice of four main chromatin states (Fig. III.4A) can thus be seen as an attempt to test the limits of the classical dichotomic picture [26,36,90] of two chromatin states, one open (euchromatin) and another one closed (heterochromatin) (Fig. III.5A).

III.2.2 Epigenetic content of the four prevalent chromatin states

The four prevalent chromatin states so identified and further labeled C1, C2, C3 and C4, were respectively found in 6572 (23.8%), 5312 (19.2%), 6603 (23.9%) and 6758 (24.4%) among the 27656 100kb loci with a defined MRT (Sect. III.4.1). Indeed, we removed from the analysis the 2411 (8.7%) loci that were not properly classified in any chromatin state. More than 90% of the loci in C1 are associated (positive enrichment) with the histone modifications H3K36me3, H3K4me3, H3K27ac and H3K79me2, the hallmarks of transcriptionally active chromatin (Fig. III.6) [6,55,166], as well as of the loci associated with RNA Polymerase II (Fig. III.7) and the RPD3-interacting protein SIN3A (Fig. III.7) as previously found in active euchromatin in *Drosophila* [50]. The majority of C1 loci are marked by H3k9me1loci consistently with the observation of higher H3K9me1 levels in active promoters [166], and also contains the histone variant H2AZ whose binding level was shown to correlate with gene activity in human [166] (Fig. III.6). C2 is notably associated with the histone modification H3K27me3 (Fig. III.6), hence corresponds to a Polycomb repressed facultative heterochromatin state [145, 166]. Out of the four main chromatin states, C3 corresponds to 100 kb loci are not enriched for any available marks. C3 can be compared to the “null" or “black" silent heterochromatin regions previously found in *Drosophila* [50, 53] and *Arabidopsis* [51] as covering a significant portion of the genome. C4 corresponds to the classic HP1-associated heterochromatin state with all of the 6603 C4 100-kb-loci containing the H3K9me3 mark and almost only that repressive mark (Fig. III.6) [145, 166].

Methylation of H3K9 is well known to be implicated in heterochromatin formation and gene silencing [55]. The fact that H3K9me1 is found almost

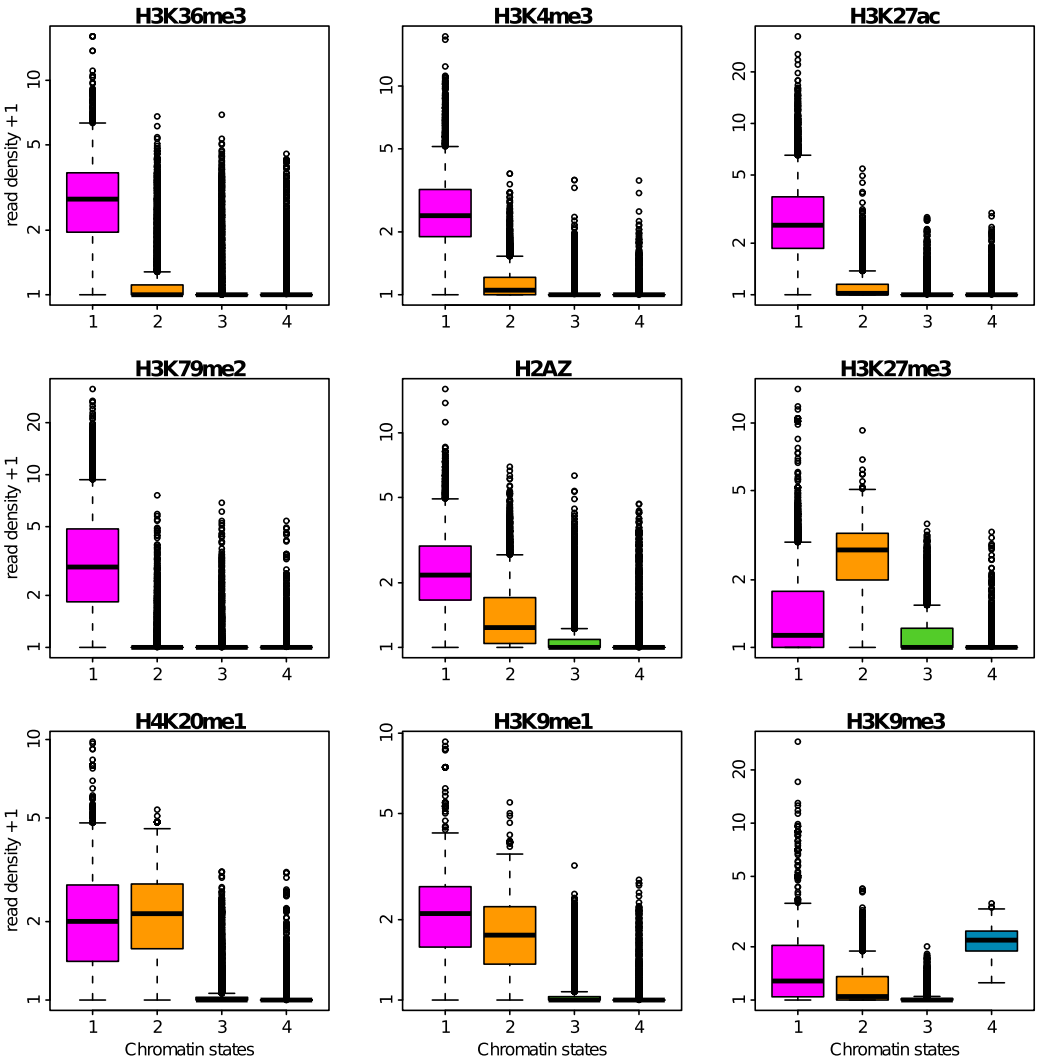

Figure III.6: Repartition of histone marks in the four chromatin states. Boxplots of the decimal logarithm of histone mark ChiP-seq read density in 100 kb non-overlapping windows per chromatin state. Same color coding as in Fig. III.4A.

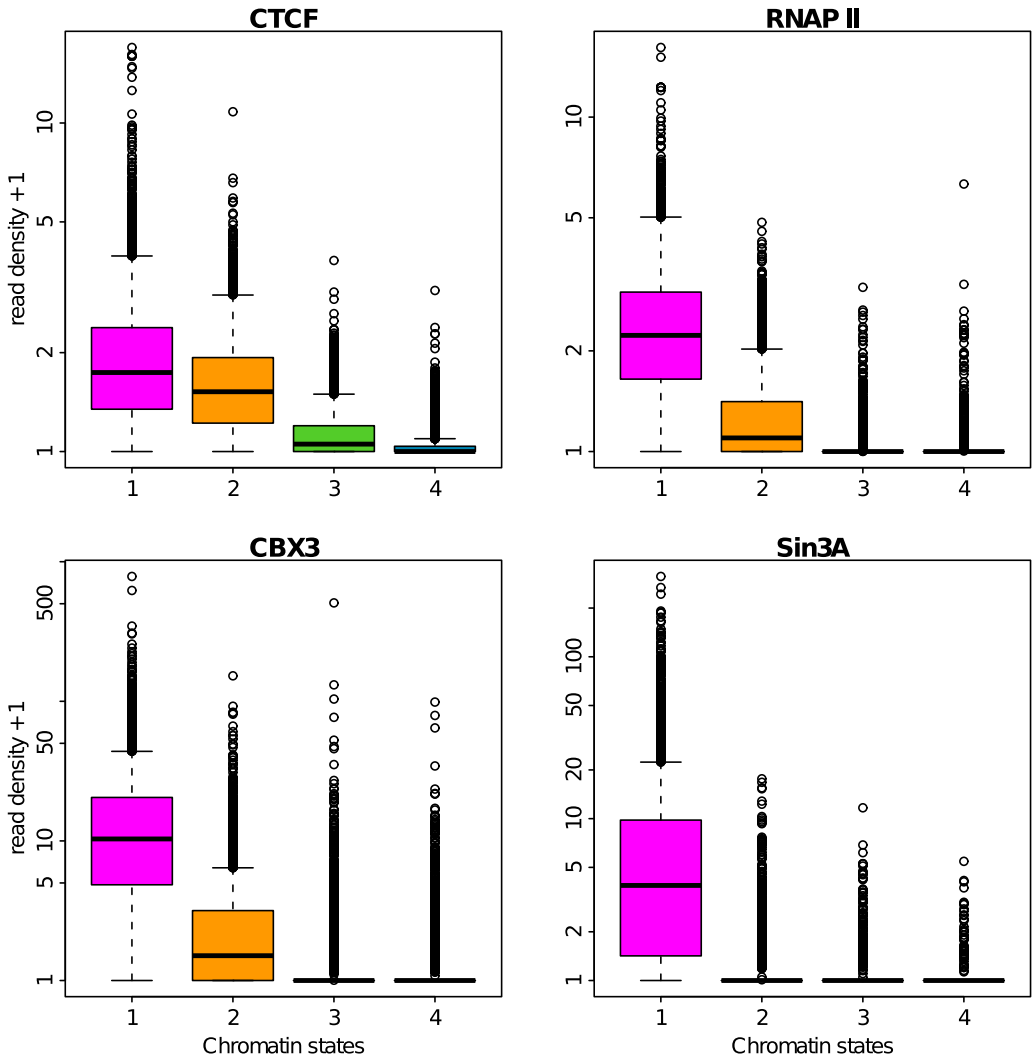

Figure III.7: Repartition of transcription factors in the four chromatin states. Boxplots of the decimal logarithm of transcription factor ChiP-seq read density in 100 kb non-overlapping windows per chromatin state. Same color coding as in Fig. III.4A.

equally in C1 and C2 and not in C4 (Fig. III.6), confirms that this epigenetic modification may also be associated with transcriptional activation [166]. H3K9me3 is found in all C4 100-kb-loci as the probable signature of its ability to anchor the heterochromatin protein HP1 at the origin of the establishment of heterochromatin. But H3K9me3 is not exclusively found in C4 loci; indeed 75% of C1 loci and 50% of C2 loci contain some H3K9me3 marks (Fig. III.6). In the transcriptionally active state C1, H3K9me3 is present in combination with all active marks which might conduct in the anchoring of the γ isoform of the HP1 protein [170–173], also called CBX3 (Fig. III.7), which was recently shown to help the splicing of multiexonic genes [174, 175] .

The insulator-binding protein CTCF is known to establish chromatin boundaries to prevent the spreading of heterochromatin into transcriptionally active regions [142, 176]. Consistent with the idea that CTCF-bound insulators prevent heterochromatin to invade genic regions, we found in good agreement with previous observation in *Drosophila* [50, 53] that CTCF is contained in C1 loci and to a slightly less extent in C2 loci (Fig. III.7).

Despite the original association of H4K20 methylation with repressive chromatin [55], H4K20me1 was recently shown to strongly correlate with gene activation [166]. In particular when combined with H3K36me3 and H2BK5me1, this mark was found at highly expressed exons near human gene 5'-ends [177]. The high level of H4K20me1 found in C1 (Fig III.6) is quite consistent with these observations. However, we observed the same level of H4K20me1 in C2 which is silent. This suggests that this mark is not uniquely linked to transcription activation. Interestingly, recent works have confirmed that PR-Set7 involved in the deposition of H4K20me1 plays an important role in the control of replication origin firing in mammalian cells [178–180].

To assess the generality of the four prevalent chromatin states, we ran the same clustering procedure on the lymphoblastoid cell line GM12878 and on a third blood cell line (Monocyte CD14+, Monocd14ro1746). The same four main chromatin states emerged in the three cell lines (Figs III.8 and III.9). Hence the chromatin organization in four chromatin states is shared by at least several somatic human cell lines.

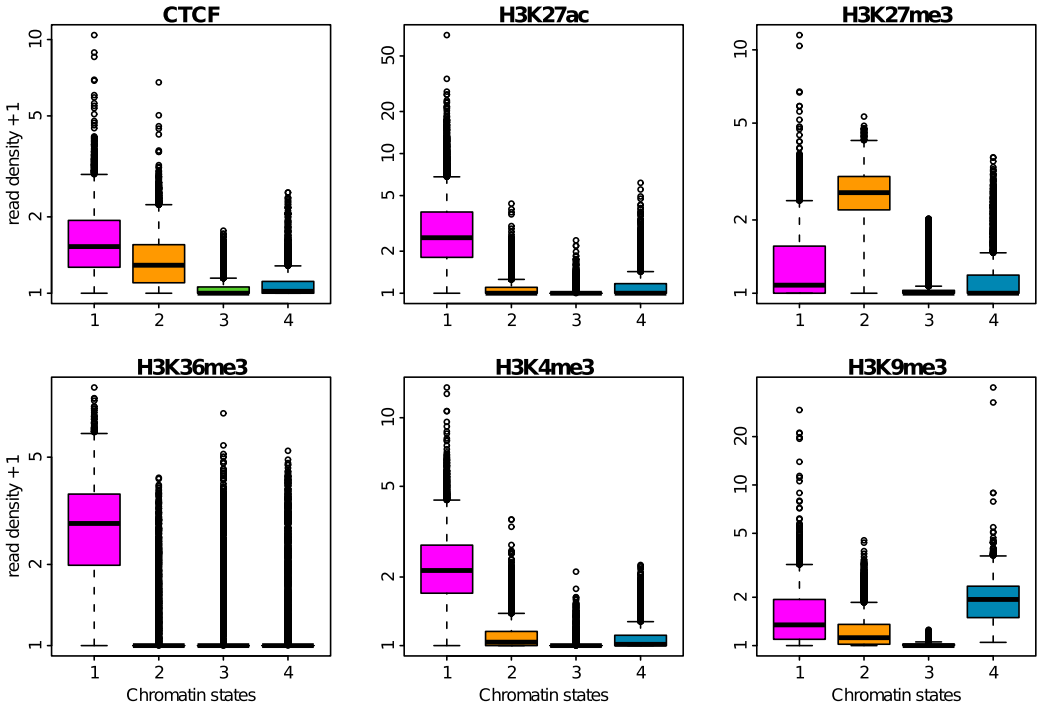

Figure III.8: Repartition of epigenetic marks in the four chromatin states for the GM12878 cell line. Boxplots of the decimal logarithm of epigenetic mark ChIP-seq read density in 100 kb non-overlapping windows per chromatin state. Same color coding as in Fig. III.4A.

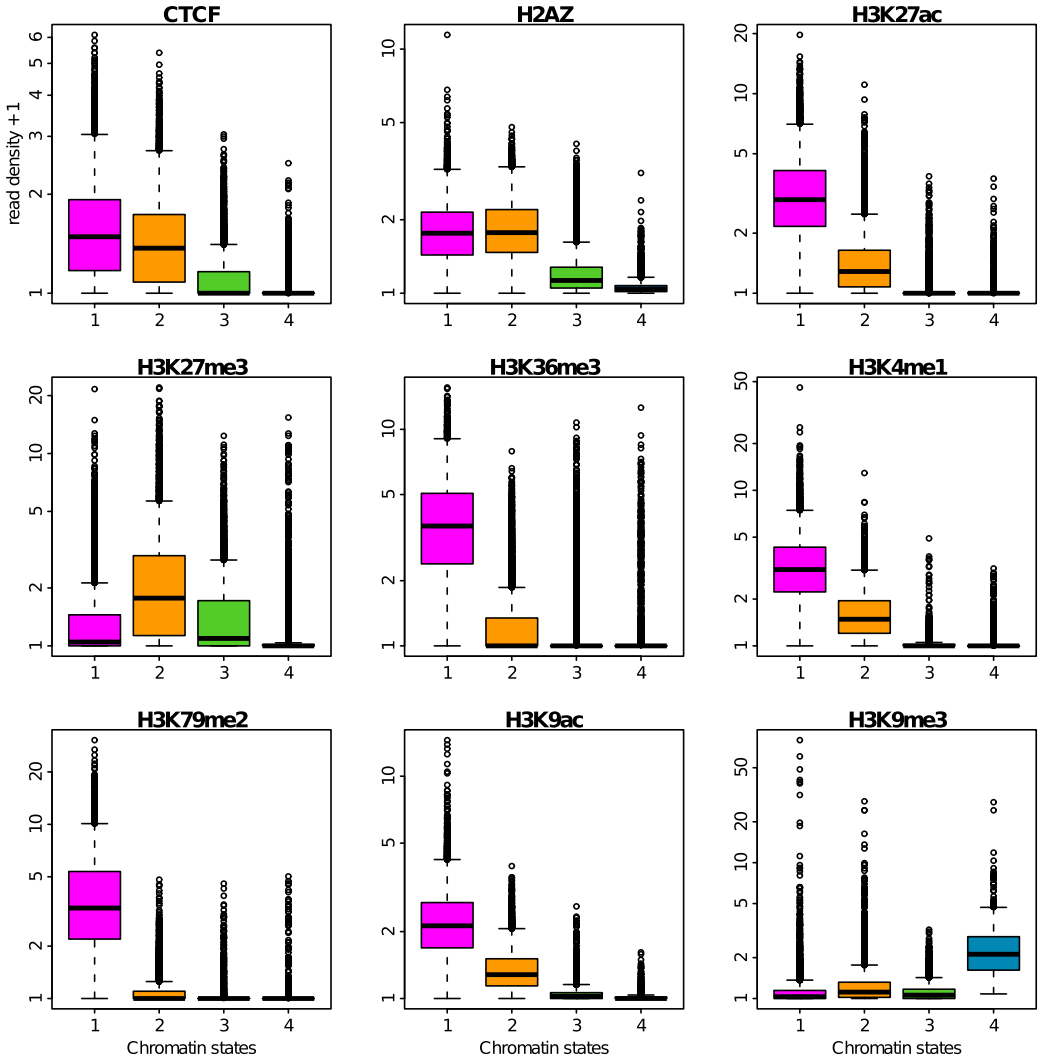

Figure III.9: Repartition of epigenetic marks in the four chromatin states for the Monocd14ro1746 cell line. Boxplots of the decimal logarithm of epigenetic mark ChIP-seq read density in 100 kb non-overlapping windows per chromatin state. Same color coding as in Fig. III.4A.

III.2.3 Chromatin states are replicated at different times during S phase

This classification into four main chromatin states of the human genome shows strong similarities with those recently reported in *Arabidopsis* [51] and *Drosophila* [50, 53] suggesting the possible existence of some simple principles of epigenetic compartimentalization of eukaryotic genomes. However, what our study reveals with respect to previous works, is a strong correlation between these chromatin states and MRT (Fig. III.10). C1, C2, C3 and C4 actually have significantly different MRT probability distribution functions (Fig. III.10A) with a clear shift from early to late replicating as evidenced by the cumulative distribution functions (Fig. III.10B). By applying a wilcoxon test to each pairs of chromatin states, we did verify that the p-value was infinitesimal. The transcriptionally active euchromatin state C1 replicates early in S phase consistent with previous analysis of open chromatin marks in human and mouse [12,14,34,36,87,89]. The Pc-repressed facultative heterochromatin state C2 is replicated slightly later in mid-S phase which corroborates the recent finding of an association of H3K27me3 with mid-replicating chromosomal domains in human fibroblast [145]. This rather clear observation contrasts with previous contradictory results concerning the existence of high correlation between late replication and this repressive chromatin mark [36, 181]. The silent unmarked chromatin state C3 replicates later than C2 but before the HP1-associated heterochromatin state C4 that replicates very late almost at the end of S phase (Fig. III.10). As previously reported in *Drosophila* [50, 88], these results confirm the existence of a strong link between epigenetic chromatin states and MRT in human. They further suggest that the epigenetically controlled chromatin structure has some impact on the normal progression of S-phase. Note that similar results were found for the GM12878 cell line (Fig. III.11) up to some slight exchange in late replication timing between C3 and C4.

III.2.4 Chromatin states are different functionally

To address the question of the gene content of these four prevalent chromatin states, we used a data set of 23818 genes that are spatially distinct (Sect. III.4.8). Some of these genes (3001) were not taken into account in our analysis because their promoter don't belong to any chromatin state. The

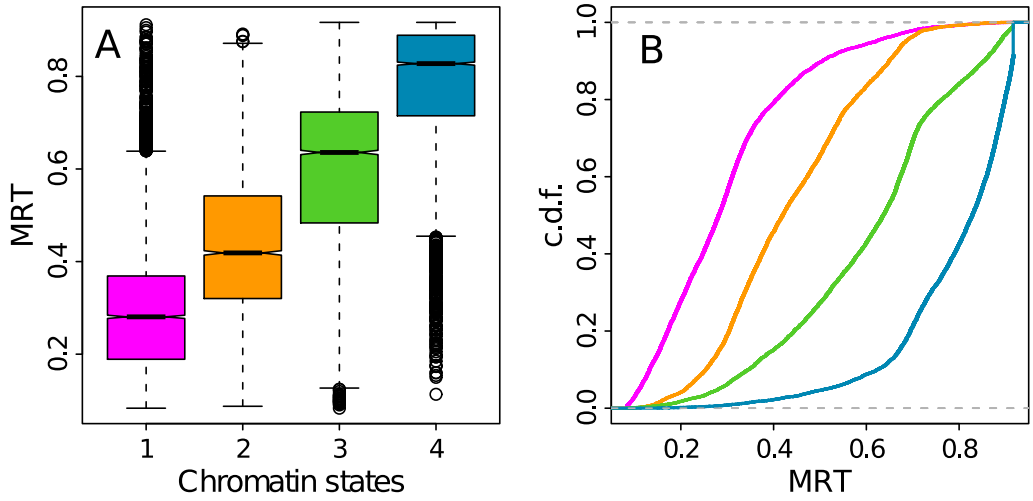

Figure III.10: MRT in the four chromatin states. (A) Boxplots of MRT computed in 100 kb non-overlapping windows per chromatin state. (B) Empirical cumulative distribution function (c.d.f.) of MRT in the four chromatin states. Same color coding as in Fig. III.4A.

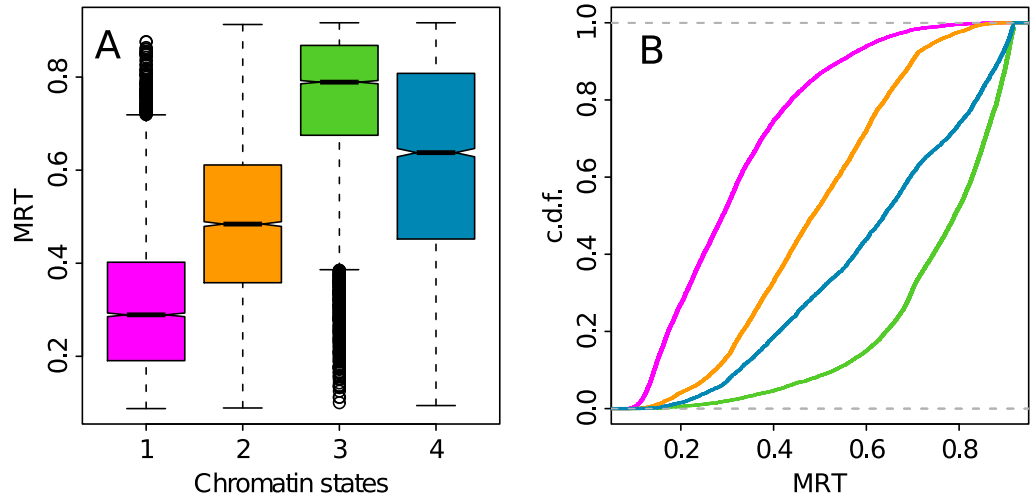

Figure III.11: MRT in the four chromatin states for the GM12878 cell line. (A) Boxplots of MRT computed in 100 kb non-overlapping windows per chromatin state. (B) Empirical cumulative distribution function (c.d.f.) of MRT in the four chromatin states. Same color coding as in Fig. III.4A.

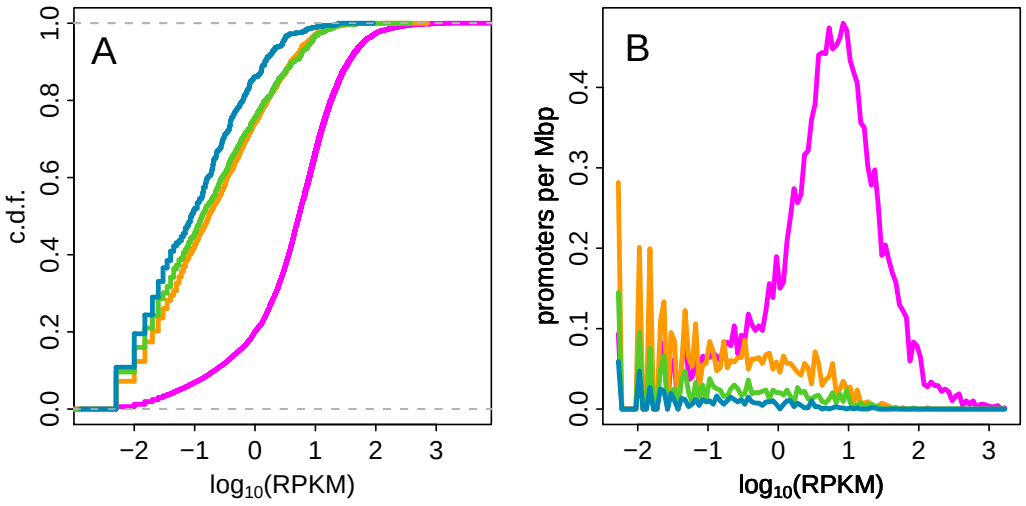

Figure III.12: Gene expression in the four chromatin states. (A) c.d.f. of gene expression (measured in $\log_{10}$(RPKM), (Eq. (III.7)) in the four chromatin states. (B) Density of promoters in the 4 chromatin states as a function of gene expression (genes were grouped into bins of width 0.05 in $\log_{10}$(RPKM) unit). Same color coding as in Fig. III.4A.

mean density of the 20817 genes that belong to one of the four chromatin states is 8.24 promoters per Mb. The only chromatin state that is highly enriched in gene promoters is the early replicating euchromatin state C1 that harbours 62.0 % of gene promoters even though it represents about 25% of the total genome coverage by the four chromatin states (Tables III.1 and III.3). The mid S facultative heterochromatin state C2 also contains a non negligible percentage (19.6%) of gene promoters that indeed corresponds to a modest density 7.7 promoters/Mb as compared to 19.1 promoter/Mb found in C1. The late replicating unmarked and constitutive heterochromatin states C3 and C4 are genuinely gene deserts with very low gene densities 4.1 promoters/Mb and 1.8 promoter/Mb respectively. The mean gene length increases gradually from C1 to C4 going from 42.5 kb to 133.1 kb (Table III.1). This discrepancy in gene length explains why the gene coverage decreases less abruptly than the promoter density, with C1 mainly genic (62.9%), C2 modestly genic (49.8%) and C3 (39.5%) and C4 (29.3%) mostly intergenic.

To investigate gene expression in chromatin states, we used a data set of 17872 genes with a valid expression value in K562 (Sect. III.4.8). Of those genes, 15869 belong to one of the chromatin states. We found that a vast majority of expressed genes with a RPKM > 1 (Eq. (III.7)) are in

Chromatin states	C1	C2	C3	C4
gene fraction (percent)	62.0	19.6	12.6	5.8
gene density per Mb	19.1	7.7	4.1	1.8
median gene length (kb)	19.0	19.0	17.8	26.1
mean gene length (kb)	42.5	59.4	83.5	133.1
gene coverage (percent)	62.9	49.8	39.5	29.3

Table III.1: Gene content in the four chromatin states. For each chromatin state, the following information is given: (i) the fraction of genes in this state in percent of the total number of genes classified in the four chromatin states, (ii) the density of genes per Mb, (iii) the median gene length in kb, (iv) the mean gene length in kb and (v) the fraction of the chromatin state covered by genes in percent. The number of genes taken into account are 12904 genes in C1, 4089 in C2, 2625 in C3 and 1199 in C4.

the early replicating euchromatin state C1 (Fig. III.12B), which confirms the link between MRT and expressed gene density previously reported in mammals [11, 12, 14, 86]. As expected, most of the genes in the facultative Pc repressed heterochromatin state C2 are non expressed. Interestingly, we found that the density of non expressed genes in C1 is equivalent to the one in C2, indicating that it is more the predominance of active genes that characterizes early replicating regions than the absence of repressed genes. This explains why the correlation between MRT and gene expression is stronger if one considers the expressed gene density ($R = 0.58$, $P < 2.10^{-16}$)) than the mean expresssion ($R = 0.24$, $P < 2.10^{-16}$) as previously observed in *Drosophila* [85]. Indeed in C1 the mean gene expression level is lowered by the presence of a non negligible set of non-expressed genes. The few genes in the heterochromatin states C3 and C4 are silent except a minority of them.

We assessed gene function on the basis of gene ontology [182]. We analyzed the genes in each chromatin states according to their biological process (Fig. III.13), component (Fig. III.14) and function (Fig. III.15) using GO SLIM annotation (Sect. III.4.11). We computed the enrichment p-value using the Hypergeometric distribution and used the odd ratio value to determine if the deviation from expected number of genes for the considered GO terms was an enrichment (odd ratio > 1) or a depletion (odd ratio < 1). As previously observed for gene expression, these GO terms provide some clear discrimination between genes in the early replicating transcriptionally active euchromatin C1 and genes in the repressed heterochromatin states C2, C3 and C4. Genes en-

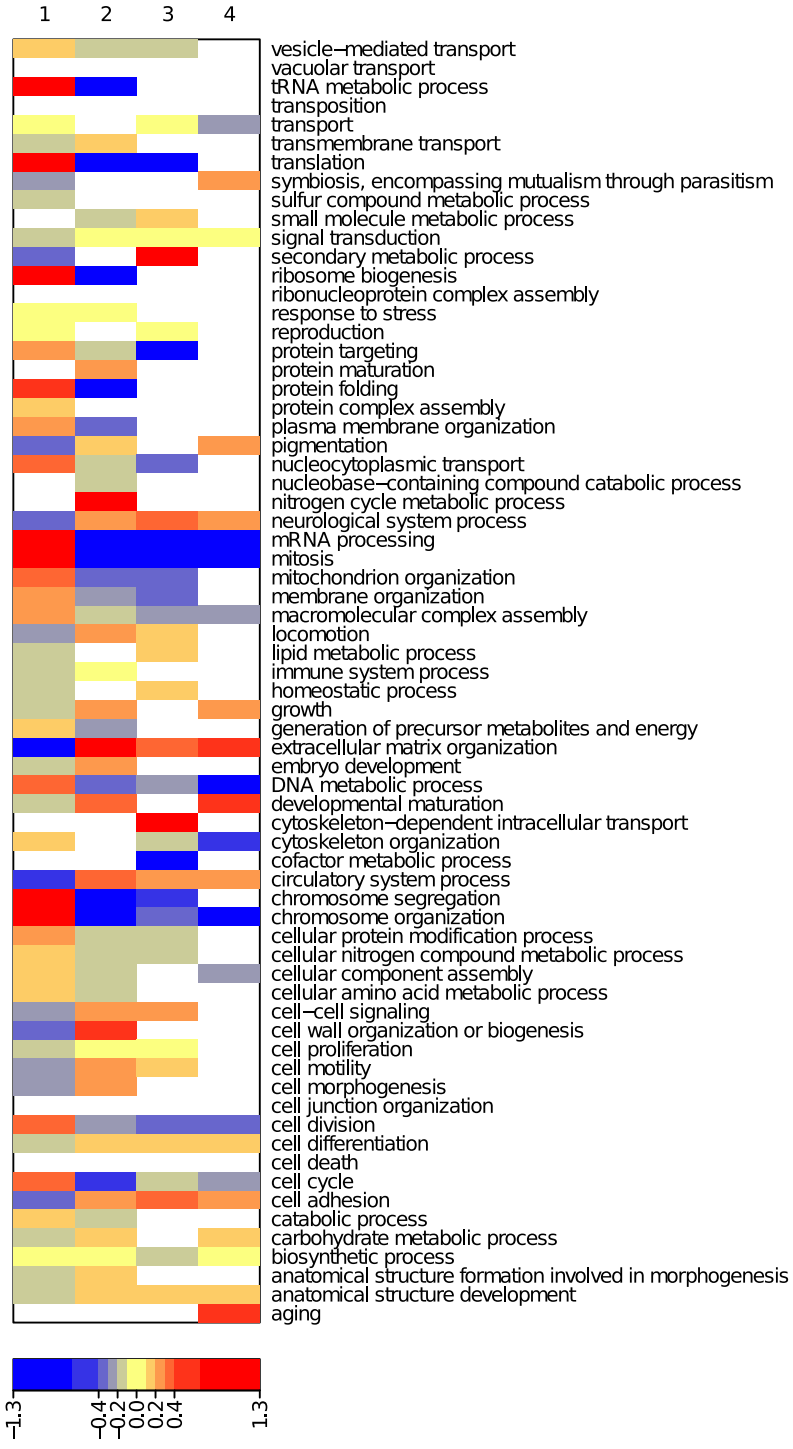

Figure III.13: GO term enrichment of the Biological Process ontology in the four chromatin states. Fisher's exact test odd ratios were computed for each GO term of the Biological Process ontology in the four chromatin states. If the test was unsignificant the corresponding cell was left blank (Sect. III.4.11) otherwise the $\log_{10}$(odd ratio) value was coded using the color map shown at the bottom.

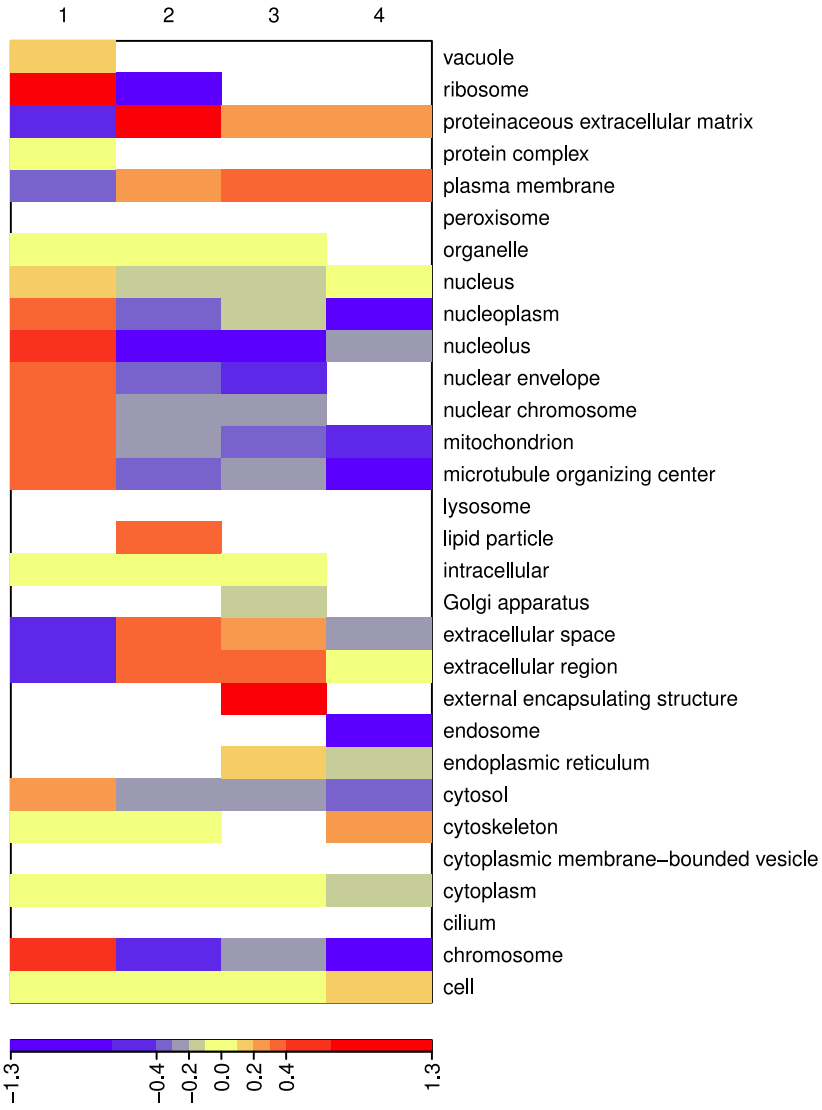

Figure III.14: GO term enrichment of the Cellular Component ontology in the four chromatin states. Same as Fig. III.13 for the Cellular Component GO term annotation.

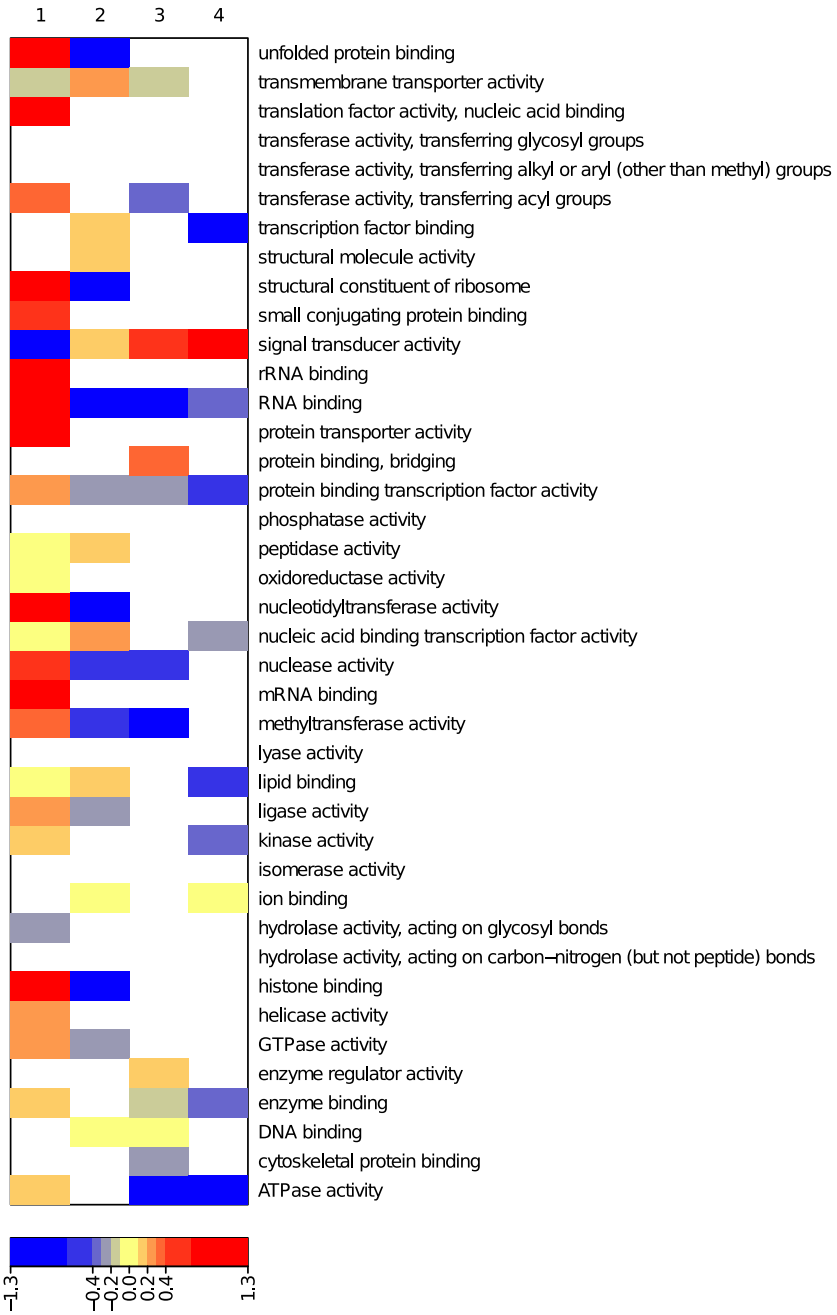

Figure III.15: GO term enrichment of the Molecular Function ontology in the four chromatin states. Same as Fig. III.13 for the Molecular Function GO term annotation.

riched in C1 are almost systematically depleted in C2, C3 and C4, whereas on the opposite, genes that are depleted in C1 are enriched in at least one if not all the heterochromatin states C2, C3 and C4. We found C1 to be enriched mainly in housekeeping genes. The highest enrichments were obtained for the following process categories: mRNA processing, translation, ribosome biogenesis, DNA metabolic process, chromosome organization and segregation, cell cycle and cell division and for the corresponding component categories: ribosome, chromosome, nucleolus, nucleoplasm, nuclear envelope, mitochondrion and microtubule organizing center. The highly depleted process categories in C1 correspond to tissue specific genes that are not expressed in the immature myeloid K562 cell line as for example neurological system process, extracellular matrix organization, cell adhesion and cell motility, or that are defficient in these cancer cells like circulating system process [183, 184].

III.2.5 Compositional content of chromatin states

Along the line of the isochore model [185], GC-rich and GC-poor regions were shown to match the cytogenic R and G bands and to correlate well with early and late replicating domains in mammals [1, 186, 187]. GC-rich regions correspond to regions of very high density of genes including the housekeeping genes and associated CpG islands. This also correspond to regions enriched in short inter-dispersed repetitive DNA elements (SINEs, Alu) [1]. In contrast, GC-poor regions are definitely poor in genes, predominantly tissue-specific genes containing rather large introns, but are relatively rich in long inter-disperse repetitive DNA elements (LINES) [1] that are significantly more abundant in these regions. Consistently, we found that the early replicating euchromatin state C1 has a GC content distribution shifted to higher values as compared to the unmarked and constitutive heterochromatin states C3 and C4 respectively (Fig. III.16A). C1 is definitely GC-rich with an mean value $\overline{\mathrm{GC}} = 44.0\%$ that is significantly higher than the genome average ($\overline{\mathrm{GC}} = 41.0\%$). On the opposite C3 and C4 are GC-poor with $\overline{\mathrm{GC}} = 39.3\%$ and 36.7%, respectively. Surprisingly, the Pc repressed facultative heterochromatin state C2 has a GC content distribution similar to the one obtained for C1 (Fig. III.16A) with $\overline{\mathrm{GC}} = 44.0\%$. This means that if a high density of early replicating and highly expressed genes implies a high GC content, the reciprocal is not true. For example, C2 loci corresponding to 18% of the genome are GC-rich (Fig. III.16A) but gene poor (Table III.1) and most of these C2 genes are silenced by Pc

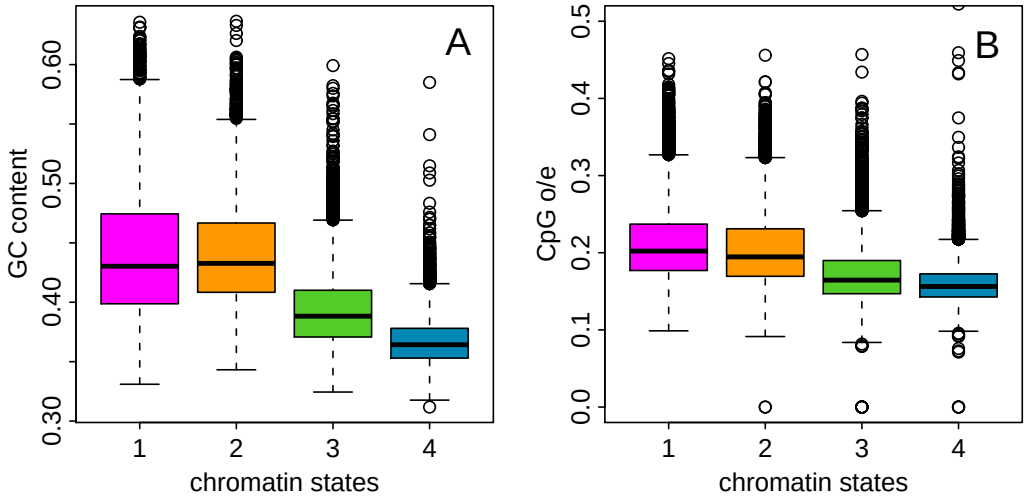

Figure III.16: Sequence composition in the four chromatin states. (A) Boxplots of GC content computed in 100 kb non-overlapping windows per chromatin state. (B) Boxplots of CpG o/e computed in 100 kb non-overlapping windows per chromatin states. Same color coding as in Fig. III.4A.

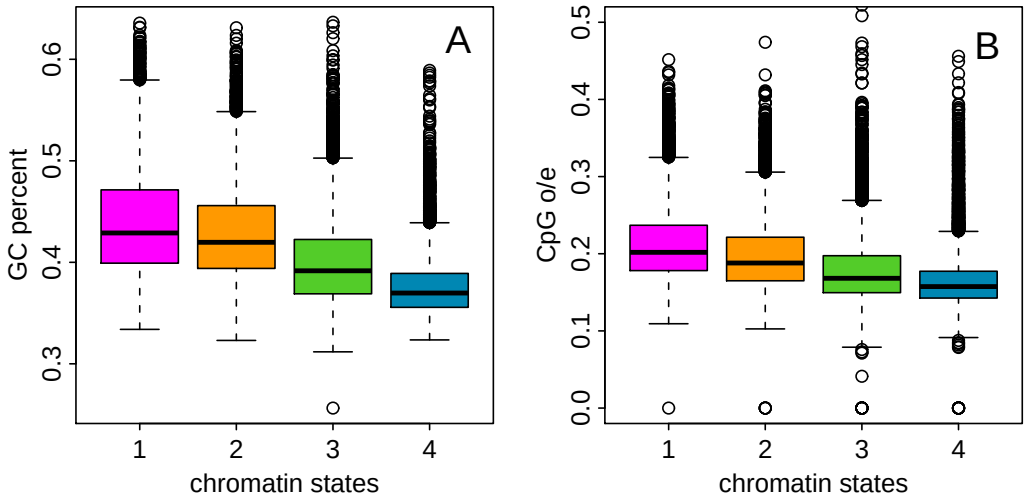

Figure III.17: Sequence composition in the four chromatin states in the Monocd14ro1746 cell line. (A) Boxplots of GC percent computed in 100 kb non-overlapping windows per chromatin state. (B) Boxplots of CpG o/e computed in 100 kb non-overlapping windows per chromatin states. Same color coding as in Fig. III.4A

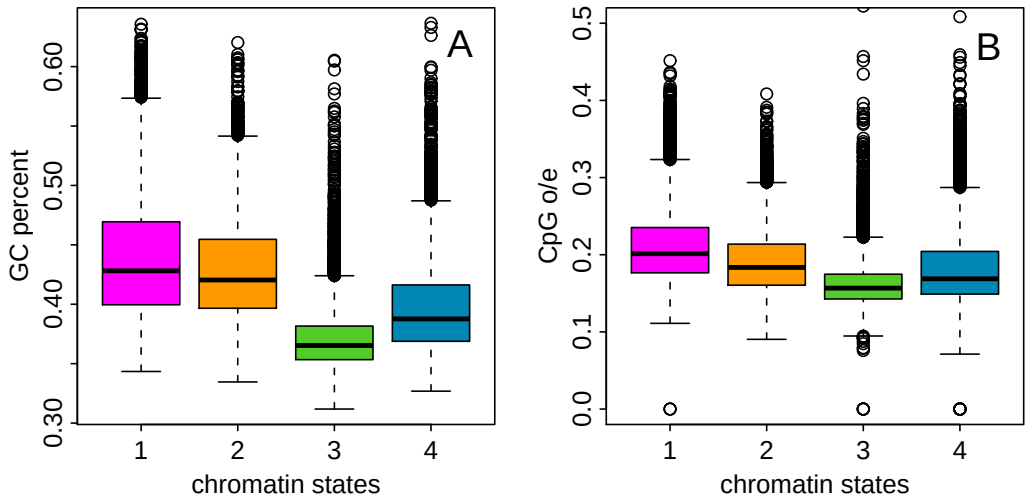

Figure III.18: Sequence composition in the four chromatin states in the GM12878 cell line. (A) Boxplots of GC percent computed in 100 kb non-overlapping windows per chromatin state. (B) Boxplots of CpG o/e computed in 100 kb non-overlapping windows per chromatin states. Same color coding as in Fig. III.4A.

proteins.

Cytosine DNA methylation is a mediator of gene silencing in repressed heterochromatin regions, while in potentially active open chromatin regions DNA is essentially unmethylated [188, 189]. Methyl-cytosines being hypermutable, prone to deamination to thymines, CpG o/e ratio (Sect. III.4.9) is commonly used as an estimator of DNA methylation, the higher this ratio, the lower the methylation [112, 190]. When computing CpG o/e after removing the CpG islands (CGIs) that are short unmethylated regions rich in CpG, in the four chromatin states, we found a significant shift of the CpG o/e pdf to smaller values when going from C1 ($\overline{\text{CpG o/e}} = 0.202$) to C2 ($\overline{\text{CpG o/e}} = 0.195$), C3 ($\overline{\text{CpG o/e}} = 0.164$) and C4 ($\overline{\text{CpG o/e}} = 0.156$) (Fig. III.16B). Thus relative to the genome average value $\overline{\text{CpG o/e}} = 0.177$, the early replicating transcriptionally active euchromatin state C1 is clearly hypomethylated. The mid-S repressed facultative heterochromatin state C2 is also, but at a lesser extent, less methylated than the entire genome. As expected the late replicating unmarked and constitutive heterochromatin states C3 and C4 are definitely methylated, the later being significantly more methylated than the entire genome. Thus the differences in CpG o/e (Fig. III.16B) and MRT (Fig. III.10A) observed in the four chromatin states C1, C2, C3 and C4, explain the significant correlation

observed genome wide between methylation and replication timing ($R = 0.402$, $P < 2.10^{-16}$)) [112].

Note that chromatin state compositional content in Monocd14ro1746 is quite the same as in K562 (Fig. III.17). In constrast, C3 and C4 in GM12878 have exchanged their GC and CpGo/e distributions (Fig. III.18). Interestingly, this phenomenon is paired with C3 becoming more late in GM12878 than C4 (Fig. III.11). This observation suggests that the genomic regions that replicate late in S phase are more likely specified by sequence features than by epigenetic features. However, the GC content cannot be the primary determinant of MRT for C1 and C2 states. Indeed the GC distributions in C1 and C2 are nearly the same (Figs III.16A, III.18A and III.17A) whereas a great discrepancy is observed in the MRT distributions (Figs. III.10, III.11 and MRT data non available).

III.2.6 Repartition of chromatin states along human chromosomes

Once mapped on the genome (Fig. III.19A,B), the four prevalent chromatin states differ not so much in the genome coverage but mainly in their number and length distribution of domains or blocks of adjacent 100-kb-loci in the same chromatin state (Table III.2 and Fig. III.19C). C1 and C2 chromatin blocks are more numerous but they are shorter with a mean length $\overline{L} = 275$ kb and 228 kb respectively. Their length pdfs do not reveal many domains larger than 1 Mb. C3 chromatin blocks are slightly less numerous and also mostly short, the larger mean length $\overline{L} = 325$ kb resulting from the existence of a few large C3 streches of several Mb length. The C4 block length pdf definitely differs from the previous ones by the presence of a fat tail. Not only the mean length $\overline{L} = 718$ kb is about three times the ones of C1, C2 blocks, but most of the C4 domains exceed 1 Mb up to 5 Mb and more, hence they are less numerous (Fig. III.19C). This observation is quite consistent with the HP1-associated classical heterochromatin spreading mechanism and its possible association with the nuclear envelope [6, 56].

When looking at the distribution of chromatin states along human chromosomes (Fig. III.19A,B), there is a clear evidence that C1, C2, C3 and C4 blocks are not distributed independently. In large regions with MRT $\lesssim 0.4$, short C1 and C2 blocks intersperse with each other, the C1s being the earliest

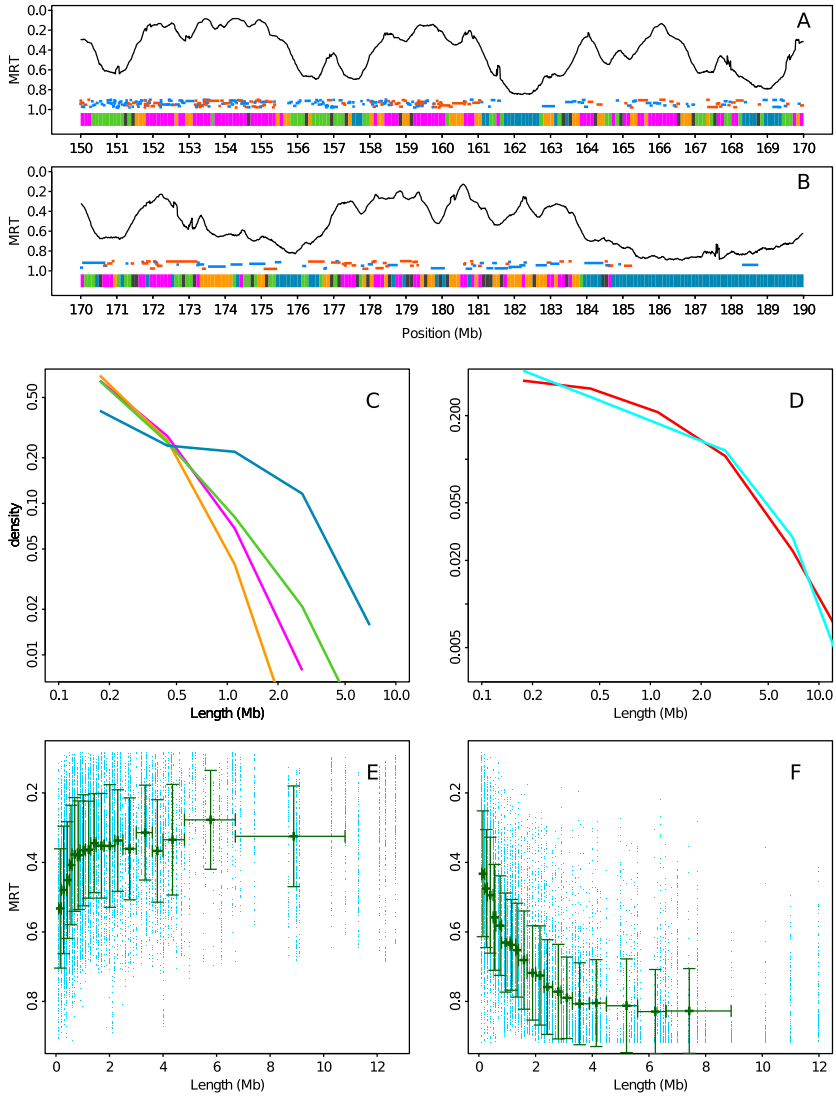

Figure III.19: Genome-wide spatial distribution of the four chromatin states. (A) MRT profile along a 20 Mb long fragment of human chromosome 1. Below the MRT profile, gene positions are indicated by a segment (blue: not expressed, orange: expressed). At the bottom of the plot, the chromatin state of each 100 kb window is represented using the same color coding as in Fig. III.4A. (B) Same as (A) for the following 20 Mb fragment of the human chromosome 1. (C) Histogram of chromatin state block length in a logarithmic representation (Sect. III.4.10). (D) Same as (C) for chromatin state blocks formed by states 1 and 2 (1+2, light red) or by states 3 and 4 (3+4, light blue). (E) MRT in chromatin state blocks (1+2) with respect to their length. Each 100 kb window in a chromatin state block is represented by a blue dot. The mean profile was obtained by (i) ordering data points according to their block length, (ii) grouping them in classes of equal number of data points and (iii) computing the average length and MRT over each class. Vertical bars represent the standard deviation. Horizontal bars represent the range of length over each class. (F) Same as (E) for chromatin state blocks (3+4).

ones (*e.g* from 158 to 161 Mb in Fig. III.19A). In a few 100kb wide regions of MRT $\backsimeq$ 0.6, C3 blocks are observed with a repressive effect (*e.g* around 156 Mb in Fig. III.19A where chromosome 1 contains a lot of olfactory receptor genes). C4 lies in very late regions MRT $\backsimeq$ 0.8 and form large uninterrupted blocks of several Mb size (*e.g* from 185 to 190 Mb in Fig. III.19A). This MRT dependent spatial organization of chromatin states prompted us to investigate neighborhood dependency between 100 kb loci. The obtained transition matrix (Table III.3) confirms that C4 loci have by far the highest probability (0.85) to have a C4 neighbor consistent with C4 blocks being much longer than the other chromatin state blocks (Table III.2 and Fig. III.19C). It also quantifies the fact that C1 loci (and in turn blocks) have a much higher probability to have a neighbor that is a C2 locus (block) than a C3 or C4 locus (block) and vice-versa. This is consistent with the fact that C1 and C2 are likely to be replicated one after each other in early and mid S phase whereas C3 and C4 are replicated much later (Fig. III.10). Consistently C4 loci (blocks) have a highest probability to have a neighbor that is a C3 locus (block) whereas C3 loci (blocks) have apparently no special preference. The spatial organization of chromatin blocks suggests that we can associate C1+C2 on one side and C3+C4 on the other side (Fig. III.5B) resulting in large-scale blocks of surprisingly very similar length distributions (Fig. III.19D) with fat tails and respective means 779 kb and 808 kb. These mega-base long C1+C2 and C3+C4 chromatin blocks would on average be replicated rather early (Fig. III.19E) and late (Fig. III.19F), respectively. Importantly, fixing the number of chromatin states to two in our PCA and cluster analysis does not result in the same dichotomic picture (Fig. III.5A). Instead we discriminate the active chromatin state C1 from a composite silent state C2+C3+C4.

Note that when using the so-computed transition matrix between chromatin states (Table III.3) to generate randomly synthetic chromosomes, we obtained very good predictions for the four chromatin state block mean lengths (Table III.2). However the corresponding sample standard deviations so predicted are significantly smaller than the ones computed for the genuine human chromosomes which is an indication that the succession of chromatin states along human chromosomes is probably governed by a more global and elaborated underlying segmentation process.

Chromatin states	C1	C2	C3	C4	C1+C2	C3+C4
total length (Mb)	674.4	533.7	641.2	676.2	1367.9	1458.3
Number	2784	2612	2305	1021	1762	1804
mean(length)	242.2	204.3	278.2	662.3	776.3	808.4
σ(length)	225.7	170.4	470.2	889.6	1171.2	1211.304
M0 mean	129	121	128	129		
M1 mean	244	204	285.7	667		
M1 σ	187.3	145.7	230.3	614.6		

Table III.2: Domain organization of chromatin states. The rows correspond to (i) the total length in Mb of each chromatin state, (ii) the number of each chromatin state domains, (iii) the mean length of each chromatin state domain in kb, (iv) the standard deviation of the length distribution for each chromatin state domain, (v) the expected length if each chromatin states were spatially independently distributed over 100-kb-loci, (vi) the expected length if 100-kb-loci chromatin state distributions are assumed to depend on their nearest neighbor and (vii) the length standard deviation given the same conditions as in (vi).

	C1	C2	C3	C4	D
	0.22	0.18	0.22	0.22	0.16
from C1	0.59	0.21	0.082	0.024	0.094
from C2	0.27	0.51	0.097	0.017	0.11
from C3	0.084	0.078	0.65	0.079	0.11
from C4	0.024	0.013	0.077	0.85	0.035
from D	0.13	0.12	0.15	0.05	0.55

Table III.3: Transition matrix between chromatin states. The first line is the probability of each chromatin state. The matrix below the first line is the Markov transition matrix between states (Sect. III.4.7). A value at the i^{th} row and the j^{th} column is the probability to find the chromatin state j in a 100 kb window next to a 100 kb window of chromatin state i. D corresponds to 100 kb windows that are not classified in any chromatin state.

III.2.7 Distribution of chromatin states inside replication timing U-domains

When concentrating our study on the 876 replication timing U-domains previously identified in K562 cells [94], we revealed some remarkable organization of the four prevalent chromatin states (Fig. III.20). The highly expressed gene rich euchromatin state C1 is found to be confined in a closed ($\lesssim$ 150kb)

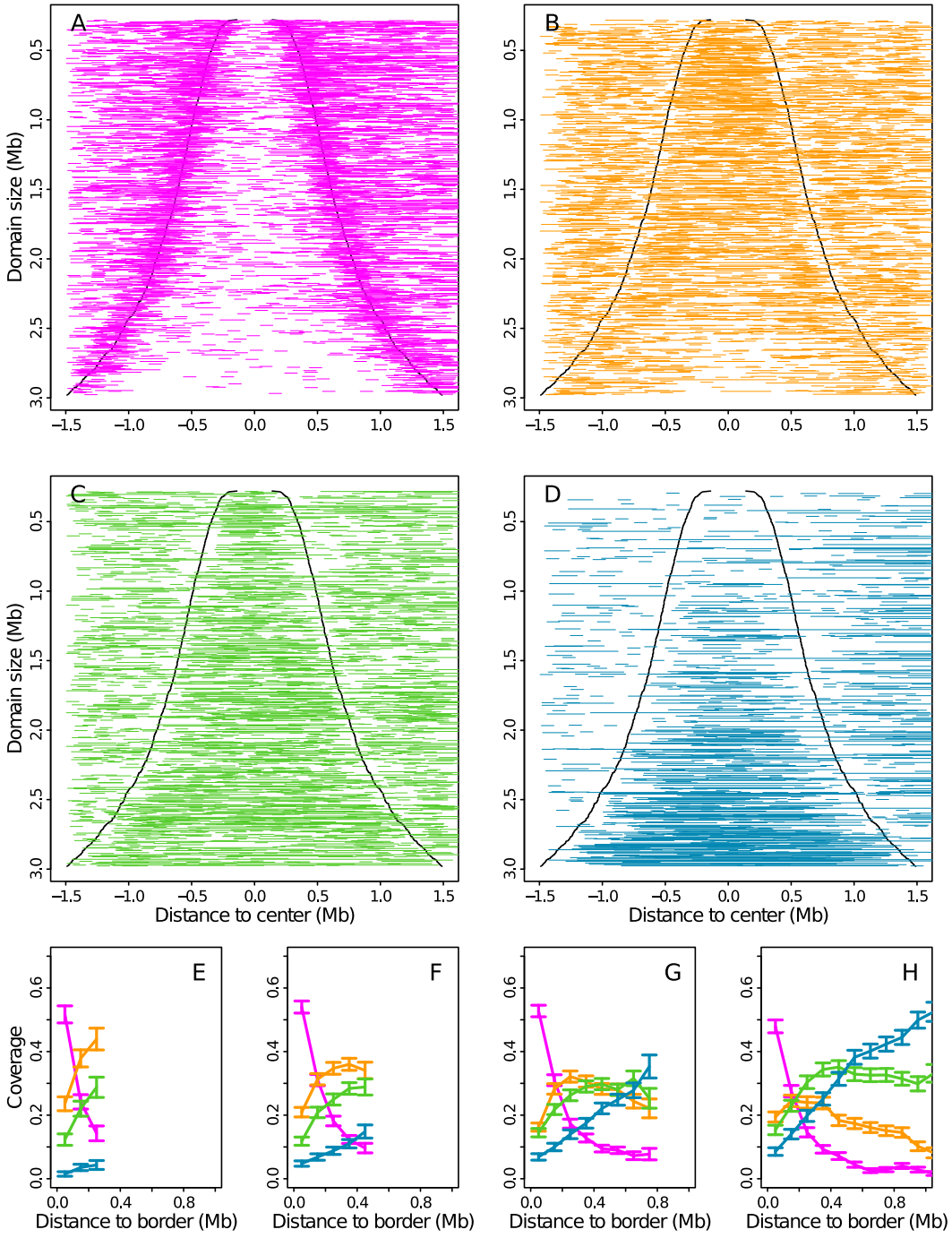

Figure III.20: Distribution of the four chromatin states inside replication timing U-domains. (A) The 876 K562 U-domains were centered and ordered vertically from the smallest (top) to the largest (bottom). All transcriptionally active chromatin state C1 100-kb-windows were represented by an horizontal segment of the corresponding length. (B) Same as (A) for the Pc repressed by chromatin state C2. (C) Same as (A) for the silent unmarked chromatin state C3. (D) Same as (A) for the HP1 heterochromatin state C4. (E) Mean coverage of chromatin state with respect to the distance to the closest U-domain border for U-domains smaller than 0.8 Mb. Error bars represent the standard deviation of the mean. (F) Same as (E) for U-domains of size between 0.8 Mb and 1.2 Mb. (G) Same as (E) for U-domains of size between 1.2 Mb and 1.8 Mb. (H) Same as (E) for U-domains of size between 1.8 Mb and 3.0 Mb. Same color coding as in Fig. III.4A.

neighborhood of the “master" replication origins that border each individual U-domains (Fig. III.20A). As confirmed on the mean occupation profiles obtained for four U-domains size categories (Fig. III.20 E, F, G, H), this confinement is independent of the U-domains size and consistent with the previous observation [94, 112] that U/N-domain borders are significantly enriched in DNase I hypersensitive sites and in insulator-binding proteins CTCF. C1 can thus be seen as specifying the early initiation zones that border U-domains and that were further shown [94, 113] to delimit topological domains on genome-wide (Hi-C) chromatin state conformation data. The Pc repressed heterochromatin state C2 is mostly found at finite distance ($\sim$ 200-300 kb) from U-domain borders as clearly seen on the largest U-domains whose centers are drastically devoided of C2 loci (Fig. III.20B,H). In small U-domains ($<$ 1.2Mb), C2 occupies in majority their centers (Fig. III.20E,F) that are replicated in mid-S phase. U-domain borders are also significantly depleted in unmarked and constitutive heterochromatin states C3 (Fig. III.20C) and C4 (Fig. III.20D), respectively. C3 is already present in the center of small U-domains (Fig. III.20E,F) and homogeneously occupies large U-domain centers (Fig. III.20G,H). C4 is significantly found in the center of U-domains that are larger than 1Mb; C4 spreads and becomes predominant when increasing the size of U-domains beyond 1.8Mb (Fig. III.20G,H). These results show that the replication “wave" starting from the early initiation zones at U-domain borders and propagating inside U-domains during S-phase with the progressive activation of secondary replication origins [93], actually corresponds to a directional path through the four prevalent chromatin states C1, C2, C3 and ultimately C4 in the largest U-domains. This gradient of chromatin structure, from active openess at U-domain borders to closeness at U-domain centers via intermediate Pc repressed and unmarked heterochromatins is likely to be a key ingredient in the long-range chromatin control of the spatio-temporal replication program that underlies the megabase-sized replication fork polarity gradients observed in about 50% of the human genome [93, 94].

III.3 Conclusion/Perspectives

In summary, this integrative analysis of epigenetic mark maps in the immature myeloid human cell line K562 has shown that the combinatorial complexity of these epigenetic data can be reduced to four prevalent chromatin states,

one transcriptionally active open euchromatin state C1 and three distinct and silent heterochromatin states, namely a Pc repressed state C2, a unmarked silent state C3 and a HP1-associated constitutive state C4. By performing this statistical study at the (low) resolution 100 kb of available genome-wide MRT data, we have found that these chromatin states actually replicate at distinct periods of the S-phase, C1 replicates early, C2 is a mid-S phase phase state whereas C3 replicates later than C2 but before C4 that replicates very late, almost at the end of S-phase. In Section III.2 are reported, for comparison, the results of a similar integrative analysis of epigenomic data in the lymphoblastoid cell line GM12878 (Figs III.8, III.11 and III.18) and in the blood cell line Monocd14ro1746 (Figs III.9, III.17), which confirm that the classification of the human epigenome in four main chromatin states likely summarizes the data in different cell types. Interestingly, these four main chromatin states display remarkable similarities with that found in different cell types in *Drosophila* [53] and *Arabidopsis* [51] at the resolution $\sim$ 1 kb of gene expression data. Suggesting the existence of simple principles of organization in metazoans as well as in plants [50–53]. When mapping these four chromatin states along the human chromosomes, our study reveals that the human genome can be segmented into megabase-sized domains of three different types with distinct spatio-temporal replication programs. In 50% of the human genome that are covered by the replication U-domains [94], the U-shape of the replication timing profile indicates that the effective replication velocity (which equals the inverse of the replication timing derivative [94, 107]) increases from U-domain borders to centers [93] as the signature of an increasing origin firing frequency during S-phase [191]. Our results (Fig. III.20) show that this acceleration of the replication wave is actually observed along a directional path through the four main chromatin states, the open euchromatin state C1 at U-domain borders successively followed by the heterochromatin states C2, C3 and C4 at the U-domain centers. To which extent this chromatin gradient influences fork progression from the “master" early initiation zones at U-domain borders and secondary origins activation inside U-domains is a key issue of current modeling [93, 131, 132, 147] of the spatio-temporal replication program in human and more generally in mammals. The complete analysis of the other half of the human genome that is complementary to U-domains is more in agreement with the traditional dichotomic picture proposed in early studies of the mouse [34, 35, 86] and human [12, 36, 90] genomes, where early and late replicating regions occur in separated compartments of open and close chro-

Chromatin states	C1	C2	C3	C4	C1+C2	C3+C4
total length(Mb)	446.3	221.8	295.2	388.8	750.6	745.5
Number	1955	1350	1216	542	1336	1031
mean(length)	228.3	164.3	242.7	717.4	561.8	723.1
σ (length)	218.2	133.6	435.0	1035.4	602.0	1275.1
M0 mean	134	115	121	130		

Table III.4: Distribution of chromatin states outside replication timing U-domains. Same as the five first lines of Table III.2 after removing the replication U-domains from the analysis.

matin, respectively. About 25% of the human genome are covered by megabase sized GC-rich (C1+C2) chromatin blocks that on average replicate early by multiple almost synchronous origins with equal proportion of forks coming from both directions (Table III.4). This absence of well-positioned origins explains that the skew has not accumulated in these gene-rich regions that were shown to be devoided of skew N-domains [102–105]. The last 25% of the human genome corresponds to megabase sized GC-poor domains of interspersed (C3+C4) heterochromatin states or of long C4 domains that on average replicate late by again multiple almost coordinated origins (Table III.4). These gene-poor regions are also devoided of skew N-domains and can be seen as the late replicating counter-part of the gene-rich (C1+C2) regions.

Extending this study to different cell types including ES, somatic and cancer cells looks very promising. By performing our integrative analysis at low (100 kb) and high (1 kb) resolutions in parallel, we should be in position to investigate the global reorganization of replication domains during differentiation (or disease) in relation to coordinated changes in chromatin state and gene expression. For example, this multivariate approach should shed a new light on the so-called replication domain "consolidation" phenomenon [34] that corresponds to the disappearance (EtoL transition) or appearance (LtoE transition) of a U-domain border during differentiation [94].The probable coordinated change in chromatin state at 100 kb resolution and the possible change at 1 kb resolution are likely to explain the possible change in gene expression. This opens new perspectives in the study of chromatin-mediated epigenetic regulation of transcription and replication in mammalian genomes in both health and disease.

III.4 Materials and Methods

III.4.1 Mean replication timing data and replication U-domain coordinates

Timing profiles for the immature myeloid cell line K562 and the lymphoblastoid cell line GM06990 were obtained from the authors [94]. The mean replication timing (MRT) is given for 27656 100 kb non-overlapping windows in hg18 coordinates. We also retrieved the coordinates of the 876 U-domains in K562 and 882 U-domains in GM06990 from the authors [94].

III.4.2 Histone marks, H2AZ, CTCF, RNAP II, Sin3A and CBX3 ChIP-seq data

For all ChIP-seq data, we downloaded data in the Encode standard format "broadpeaks" (`http://genome.ucsc.edu/FAQ/FAQformat.html`). Broadpeaks format is a table of significantly enriched genomic intervals. Most of the data correspond to the release 3 (August 2012) of the Broad histone track. We downloaded the tables from:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/`

The CBX3 and Sin3A data corresponds to the release 3 (September 2012) of the HAIB TFBS track. Tables were downloaded from the UCSC from:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/`

For the K562 cell line, we downloaded the broadpeak tables for the following antibodies: CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me3, H3K9me3, RNAP ll, H2AZ, H3K79me2, H3K9me1, H4K20me1, CBX3, Sin3A. For the GM12878 cell line, we downloaded: CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me3, H3K9me3. For the Monocd14ro1746 cell line we downloaded: CTCF, H2AZ, H3K27ac, H3K27me3, H3K36me3, H3K4me3, H3K79me2, H3K9ac, H3K9me3. Genomic intervals were then mapped back to hg18 using LiftOver.

III.4.3 Epigenetic profile computation at 100 kb resolution

For each ChIP-seq data, we computed a profile at the 100 kb resolution for the 27656 non-overlapping windows for which MRT is defined. The read density for one antibody in a window is the number of reads in this window that fall in significantly enriched intervals normalized by the window length.

III.4.4 Rank transformation and Spearman correlation matrix

All statistical computations were performed using the R software (`http://www.r-project.org/`).

In order to compute the Spearman correlation matrix, the epigenetic profiles at 100 kb resolution were transformed with the R function *rank* with option *ties.method=max*. Then we computed the Pearson correlation matrix on the transformed dataset. To reorder the matrix in Fig. III.1, we computed the Spearman correlation distance $dSCor$ as:

$$dSCor(X, Y) = 1 - SCor(X, Y), \tag{III.1}$$

where $SCor$ is the spearman correlation. Then, a dendrogram was computed using the R function *hclust* with option *method=average* and with $dSCor$ as dissimilarity.

III.4.5 Principal component analysis

Principal component analysis was performed on the rank transformed dataset using the function *dudi.pca* from the R package *ade4* (see `http://pbil.univ-lyon1.fr/ADE-4` and Ref. [167]) with the option *scale=TRUE* (*i.e.* each variable is centered and normalized before the PCA computation). The first three components were retained which accounts for 76% of the dataset variance (Fig. III.2), and clustering was performed in this 3D space.

III.4.6 Clustering strategy

We used Clara algorithm [153] which is an optimization of k-means for large data set. We used the *clara* function implemented in the R package *cluster*. The options were set to: *stand=FALSE, sampsize=500, samples=20, metric=euclidean*.

To assess the number of clusters, we used the pooled within-cluster sum of squares around the cluster mean. Suppose that the data set of size n is divided in k clusters $C_1, C_2, \ldots, C_k$. Let d(x,y) be the euclidean distance between the points x and y. Let $\overline{x}_i$ be the mean of the i^{th} cluster, then the within-cluster sum of squares for this cluster is:

$$w_i = \sum_{x_j \in C_i} d^2(\overline{x}_i, x_j). \tag{III.2}$$

The pooled within-sum of squares for the k clusters is:

$$W_k = \sum_{i=1}^{k} w_i. \tag{III.3}$$

The pooled within-cluster sum of squares necessarily decreases with the number of clusters. A good choice for the number of clusters is the critical point where some clear crossover is observed from a fast decrease of W_k at small k values to a weak decrease of W_k at large k values. This means that, after this critical point, no much information is gained by adding a new cluster. In our analysis this crossover occurs for k=4 clusters (Fig. III.4B).

We also used the Gap statistic [169] which is defined by :

$$Gap_n(k) = E_n(\ln(W_k)) - \ln(W_k). \tag{III.4}$$

$E_n(\ln(W_k))$ is the expected value of $\ln(W_k)$ for a sample of size n drawn from a proper reference distribution. We choose, as a reference, a uniform distribution over the range of the observed data. A good choice for the number of clusters is a value of k so that W_k is much smaller than the expected W_k from a random distribution (*i.e.* a high value of $Gap_n(k)$). Four clusters is also a reasonable choice according to the gap statistic index computed with R package *clusterSim* (Fig. III.4C).

Poorly clustered data points were removed from the set of chromatin states. The silhouette value [168] is a way to quantify how well a point is clustered.

Definition 1 *Given a particular clustering, $C_1, C_2, \ldots, C_k$, of the data in k clusters, let i be a data point and $d(i, C_j)$ the average distance of the data point i to the members of the cluster C_j. Let i be a member of cluster C_c and*

$$a_i = d(i, C_c), \quad b_i = min_{j \neq c}(d(i, C_j)). \tag{III.5}$$

The silhouette value of the data point i is defined as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}. \tag{III.6}$$

A silhouette value below 0 means that the data point is actually closer in average to the points from another cluster than to the one it has been assigned to. Points with a negative silhouette value are border line allocations. We decided to remove those points from the set of identified chromatin states. Hence chromatin states are groups (clusters) with homogeneous epigenetic features. 91% of all 100 kb non-overlapping windows of the human genome were assigned to one of the four chromatin states C1, C2, C3 or C4.

III.4.7 Markov transition matrix estimation

The number of transitions from i to j, n_{ij}, is the number of 100 kb windows of state i contiguous to a window of state j (the sense or antisense orientation is not taken in account). Let n_i be the number of windows in chromatin state i. The conditional probability of a transition from i to j given i is $\frac{n_{ij}}{n_i}$.

III.4.8 Annotation and Expression data

As human gene coordinates, we used the UCSC Known Genes table. When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates corresponded to the union of all the overlapping gene coordinates, resulting in 23818 distinct genes.

Expression data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC). To construct our expression data set, we used RefSeq Genes track as human gene coordinates. Genes with alternative splicing were merged into one transcript by taking the union of exons. Hence the TSS was placed at the beginning of the first exon. We obtained a table of 23329 genes. We downloaded expression valuess from the release 2 of Caltech RNA-seq track (ENCODE project at UCSC):

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeCaltechRnaSeq/

Expression for one transcript is given in reads per kilobase of exon model per million mapped reads (RPKM) [10]. RPKM is defined as:

$$R = \frac{10^9 C}{NL}, \tag{III.7}$$

where C is the number of mappable reads that fall into gene exons (union of exons for genes with alternative splicing), N is the total number of mappable reads in the experiment, and L is the total length of the exons in base pairs. We associated 17872 genes with a valid RPKM value in K562.

III.4.9 CpG o/e computation and GC content

CpG observed/expected ratio (CpG o/e) was computed as $\frac{n_{CpG}}{L-l} \times \frac{L^2}{n_C n_G}$, where n_C, n_G and n_{CpG} are the numbers of C, G and dinucleotides CG, respectively, counted along the sequence, L is the number of nonmasked nucleotides and l is the number of masked nucleotide gaps plus one, *i.e.* L-l is the number of dinucleotide sites. The CpG o/e was computed over the sequence after masking annotated CGIs. The GC content was computed on the native sequence.

III.4.10 Chromatin state blocks

We detected contiguous windows of the same chromatin state (C1 to C4). We then kept the coordinates of the blocks of contiguous windows. To form chromatin state blocks of states (1+2), we merely detected contiguous windows of state 1 or 2. The same procedure was applied to define chromatin blocks of states (3+4). For chromatin blocks (1+2) and (3+4), we authorized the inclusion of isolated windows which don't belong to any chromatin state so to not disrupt very long blocks.

III.4.11 GO term enrichment

Each gene name of our annotation dataset was associated to several GO terms from GO SLIM (high level GO terms) using the online mapper: http://go.princeton.edu/cgi-bin/GOTermMapper. Then for each chromatin state

(C1 to C4), the number of occurrences of each GO term was determined by the number of promoters belonging to that state and associated to this GO term. The enrichment for each GO term in each cluster was tested using Fisher's exact test. We applied a procedure to control the false discovery rate (FDR) as described in [192]. The upper limit of the FDR was fixed to 20%. After detecting significant deviation from a random repartition of GO term occurrences, we used the odd ratio value to determine if the deviation was an enrichment (odd ratio > 1) or a depletion (odd ratio < 1).

Chapter IV

Epigenetic regulation of the human genome: coherence between promoter activity and large-scale chromatin environment

In Chapter III, we have analyzed the genome-wide distributions of thirteen epigenetic marks in the human cell line K562 at 100 kb resolution of Mean Replication Timing (MRT) data. Using classical clustering techniques, we have shown that the combinatorial complexity of these epigenetic data can be reduced to four predominant chromatin states that replicate at different periods of the S-phase. In this Chapter, we extend our integrative analysis of epigenetic data in the K562 human cell line to a much finer scale by focusing on gene promoters ($\pm$ 3 kb around transcription start sites). We show that these promoters can similarly be classified into four main chromatin states: P1 regroups all the marks of transcriptionally active chromatin and corresponds to CpG rich promoters of highly expressed genes; P2 is notably associated with the histone modification H3K27me3 that is the mark of a polycomb repressed chromatin state; P3 corresponds to promoters that are not enriched for any available marks as the signature of a "null" or "black" silent heterochromatin state and P4 characterizes the few gene promoters that contain only the constitutive heterochromatin histone modification H3K9me3. Results reported in

this chapter are published in [193].

IV.1 Introduction

Multivariate statistical analyses of epigenetic data sets in human have revealed that distinct epigenetic modifications often exist in a well-defined combinations corresponding to different genomic elements like promoters, enhancers, exons, repeat sequences and/or to distinct modes of regulation of gene expression such as actively transcribed, silenced and poised [48, 162, 163, 177, 194]. In Chapter III, with the aim at quantifying the influence of epigenetic modifications on replication timing, we have used principal component analysis (PCA) and classical clustering method to analyze thirteen epigenetic mark maps in the K562 human cell line at the 100-kb-resolution of MRT data. This study reveals that the huge combinatorial epigenetic complexity can in fact be reduced to a rather small number of predominant chromatin states that interestingly share strong similarities with the ones previously found in *Arabidopsis thaliana* [51], *Caenorhabditis elegans* [52] and *Drosophila* [50,53]. These four main chromatin states were further shown to correlate with MRT, namely from early to late replicating, a transcriptionally active euchromatin state (C1) enriched in insulator binding protein CTCF, a polycomb repressed facultative heterochromatin state (C2), a silent heterochromatin state (C3) not enriched in any available marks and a HP1-associated heterochromatin state (C4).

In this Chapter, our goal is to extend our integrative analysis of epigenetic data in the K562 human cell line from the 100 kb scale of MRT data to a few kb scale characteristic of gene promoters as previously performed in plants [51], worm [52] and fly [50, 53]. First, we will perform a combinatorial analysis of chromatin marks in K562 around gene promoters and describe the epigenetic content of the four prevalent chromatin states. Then, we will study the coherence between promoter activity, as characterized by their "small-scale" chromatin state, and the "large-scale" chromatin environment (namely the C1, C2, C3 and C4 chromatin states found in Chapter III). In this comparative analysis we emphasize the expected as well as the unexpected importance of gene density on the observed relationship between these two scales characterizing transcription and replication data respectively. We will also investigate the spatial distribution of these promoter chromatin states inside the three types of replication domains defined in our previous work [154], namely the

50% of the human genome paved by MRT U-domains, the 25% covered by early replicating GC-rich (C1+C2) chromatin blocks and the 25% covered by late replicating, GC-poor (C3+C4 or long C4) heterochromatin blocks. We conclude, in the final section, by discussing some perspectives for further studies in different cell types, in other mammalian genomes in both health and disease.

IV.2 Combinatorial analysis of chromatin marks at human gene promoters

IV.2.1 Fine-scale analysis of chromatin marks combinatorial complexity

Mammalian promoter regions are well known to vary significantly in their positional relationships to genes [6, 55]. The DNA sequence proximal to the transcriptional start site (TSS) of a gene is commonly regarded as a proxy region where the study of chromatin marks is likely to provide new insights into the regulatory state of promoters and genes. Here we investigate relationships between the genome-wide distributions of eight histone modifications, one histone variant and four DNA binding proteins in the myelogenous leukemia human cell line K562 around ($\pm$3 kb) the 17872 gene TSS with a valid RPKM (Eq. (IV.1)). In Fig. IV.1 is shown a heat map representing the Spearman correlation matrix between epigenetic marks after having reorganized raws and columns with a hierarchical clustering algorithm based on the Spearman correlation distance [Eq. (IV.2)]. All the epigenetic marks that are known to be involved in transcription positive regulation, namely H4K20me1, H3K9me1, H3K4me3, H3K27ac, H3K79me2, RNAPII, H3K36me3, CBX3, H2AZ, together with the transcription factors CTCF and sin3A, form a block in the correlation matrix, meaning that they are all significantly correlated with each other [6, 194]. The maximum correlation is obtained between the two active promoter marks H3K4me3 and H3K27ac. Note also the preferential correlation between H4K20me1 and H3K9me1 consistent with previous observations of some enrichment of these marks in promoter or coding regions of active genes [166, 195–197], with further evidence of significant colocalization [198]. However there are mainly two lines that stand out from the block

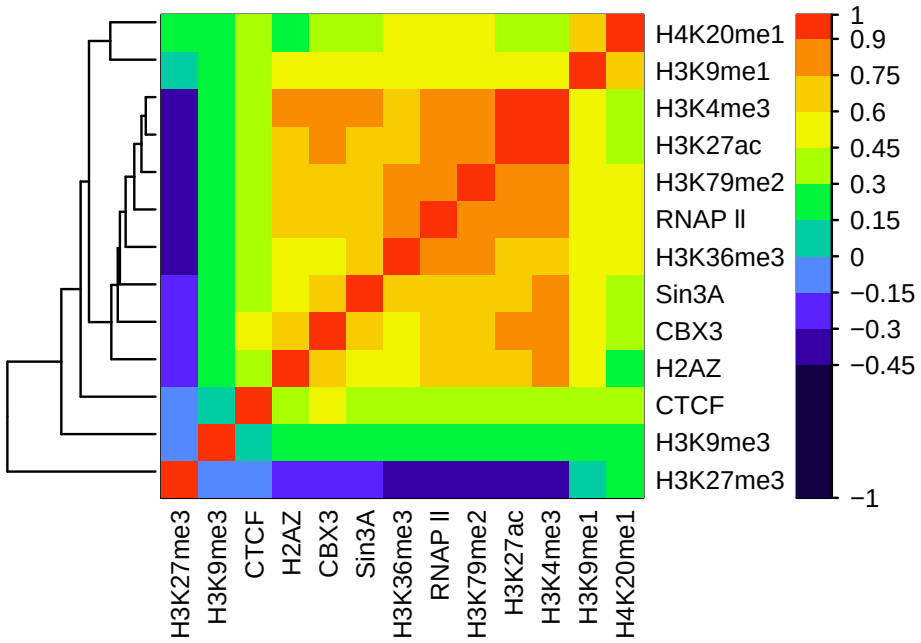

Figure IV.1: Spearman correlation matrix between epigenetic marks. For each pair of variables, we computed the Spearman correlation over 6 kb windows centered on human gene TSSs. Spearman correlation value is color coded using the color map shown on the right. Lines for the thirteen epigenetic marks were reorganized by a hierarchical clustering using Spearman correlation distances as illustrated by the dendrogram on the left of the heat map. This ordering implies that highly correlated epigenetic marks are close to each other.

of active marks in the hierarchical clustering dendrogram in Fig. IV.1. One of these lines corresponds to the polycomb (Pc) associated repressive chromatin marks H3K27me3 characteristics of the so-called facultative heterochromatin [145, 166]. This is the only mark that anti-correlates with most of the active marks except H4K20me1. The other line corresponds to H3K9me3, commonly considered as a repressive chromatin mark associated with the heterochromatin protein 1 (HP1) known as a major actor in constitutive heterochromatin formation [145, 166]. Surprisingly H3K9me3 is found to moderately correlate with all active marks. This confirms previous observations that this epigenetic modification may also be associated with transcriptional activation. When H3K9me3 is present in the promoter region in combination with all active marks, this may conduct in the anchoring of the γ isoform of the HP1 protein [170–173], also called CBX3, which was recently shown to help

the splicing of multi-exonic genes [174, 175].

IV.2.2 Principal promoter chromatin states

To objectively identify the prevalent combinatorial patterns of the thirteen chromatin marks at human gene promoters, we have performed a PCA [167] to reduce the dimensionality of the data (Sect. IV.6.5). As shown in Fig. IV.2, the first three principal components sum up 74% of the total data variance (Fig. IV.2B,C). By projecting the 6 kb promoter loci on the (PC1, PC2), (PC3, PC2), (PC1, PC3) and (PC4, PC5) planes (Fig. IV.2A), it is clear that most of the population is confined in the (PC1, PC2) plane. In this very dense plane, loci mainly lie along two straight lines with a very high density of loci at the intersection of these two lines. A rather wide diluted mode is observed parallel to the PC1 axis, whereas a more populated mode is concentrated along a line parallel to PC2. Furthermore, a simple inspection of the projections on the planes (PC3, PC2) and (PC1, PC3) in Fig. IV.2A confirms that loci out of the (PC1, PC2) plane are rather scarce (less than 5% of the human gene promoters). This has led us to phemomenologically define four main promoter chromatin states in the 3D-space defined by Eqs (IV.3) to (IV.6). When labeling each of these four promoter chromatin states with a color, namely P1 (pink), P2 (orange), P3 (green) and P4 (blue), we obtain the density contour plots shown in Fig. IV.3. Among the first three chromatin states that are confined in the (PC1, PC2) plane, P1 is by far the most populated state $N = 9643$ (54.4%) promoter loci as compared to P2 with $N = 3149$ (17.8%) and P3 with $N = 4252$ (24.0%). The fourth promoter chromatin state P4 is the only one that lie outside the (PC1, PC2) plane along a direction parallel to the PC3-axis (Eqs (IV.3) and Fig. IV.3B). This state contains only $N = 679$ (3.8%) promoter loci, which is dramatically less than the P1, P2 and P3 populations. Since, as we will see in the next sections, P4 will turn out to be a relevant and epigenetically meaningful chromatin state, the fact that classical clustering algorithms similar to k-means would have missed this very poorly populated state (see [199] for the limitations of these clustering methods) justifies, *a posteriori*, our phemomenological clustering in the four chromatin states defined by Eqs (IV.3) to (IV.6).

Remark: When using the Clara clustering algorithm [153] with the number of clusters fixed to four, we miss the chromatin state P4 that is then included in

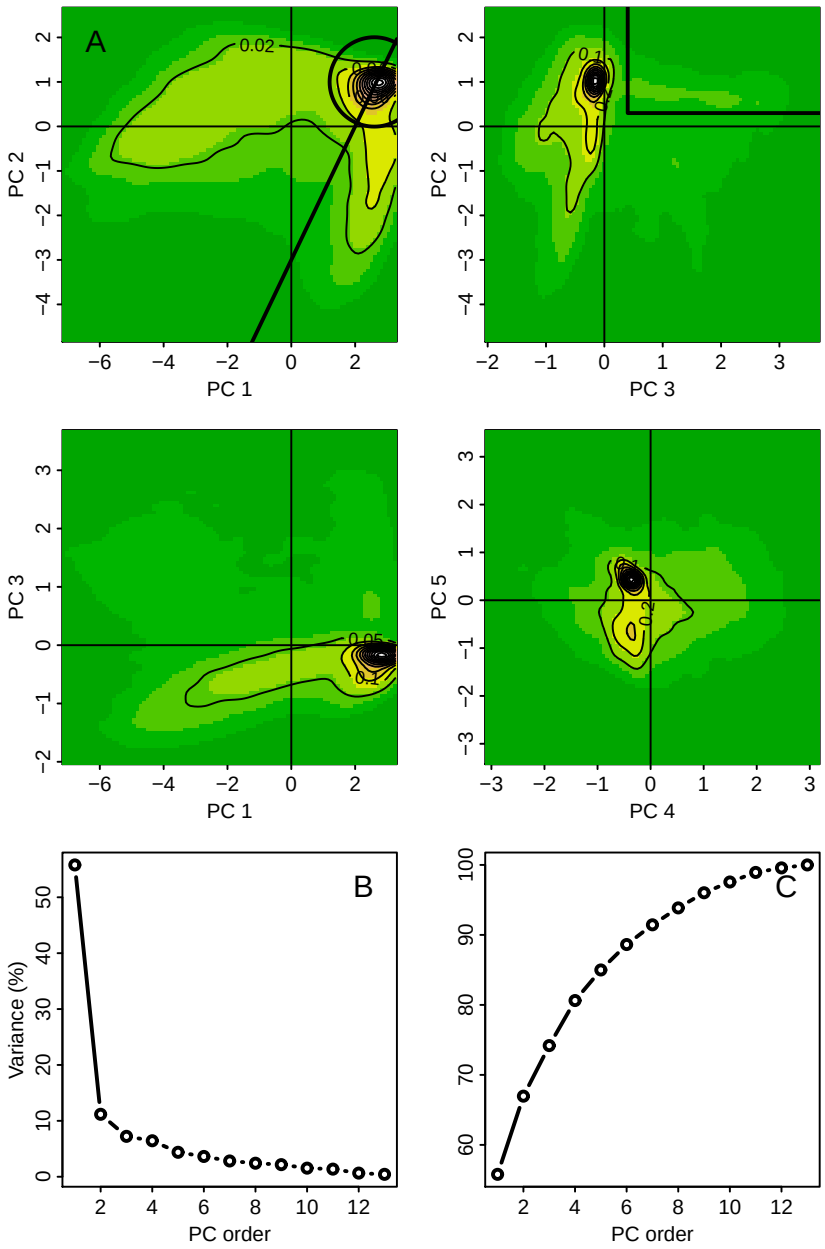

Figure IV.2: (A) Two-dimensional (2D) projections of the 6 kb promoter data points on the planes defined by (top left) the first (PC1) and second (PC2) principal components, (top right) PC3 and PC2, (bottom left) PC1 and PC3 and (bottom right) PC4 and PC5. The density values are indicated by a color code (white: high density, yellow: moderate density, green: low density) and a contour plot. Densities are computed with a kernel density estimator. The thick solid line are the boundaries that separate promoter chromatin states P1, P2, P3 and P4 in the the 3D space (PC1, PC2, PC3) as defined in Eqs (IV.3) to (IV.6). (B) Percentage of variance accounted by the first thirteen principal components ordered according to their corresponding variance (eigenvalues). (C) Cumulative variance.

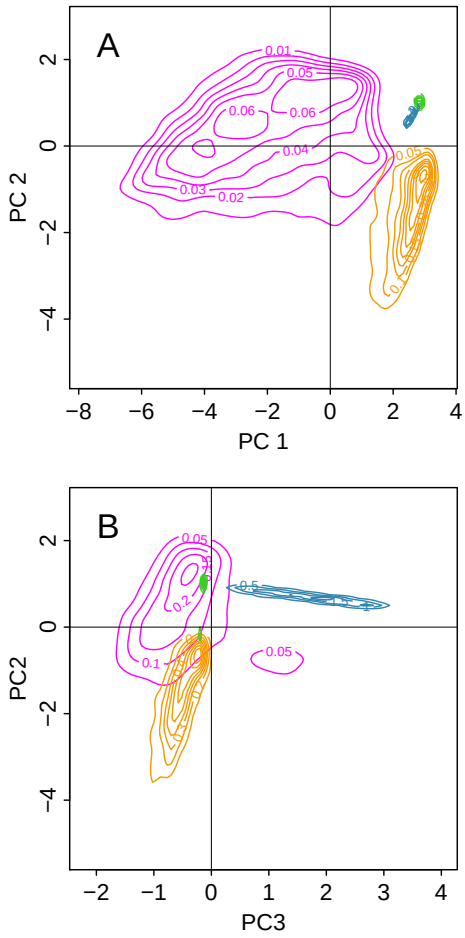

Figure IV.3: Contour plots of the densities of the four prevalent promoter groups P1 (pink): activated promoters, P2 (orange): Pc repressed promoters, P3 (green) unmarked promoters and P4 (blue): HP1 repressed promoters. The clustering in the 3D space generated by the first three principal component PC1, PC2 and PC3 is defined in Eqs (IV.3) to (IV.6) and illustrated in Fig. IV.2A. (A) 2D-projection on the plane defined by PC1 and PC2; (B) 2D-projection on the plane defined by PC3 and PC2.

P3, whereas the most populated chromatin state P1 is splitted in two states. Indeed, Eqs (IV.4), (IV.5) and (IV.6) that respectively define the chromatin states P3, P2 and P1 as mainly confined in the (PC1, PC2) plane, are inspired from the partitioning provided by the Clara algorithm. Let us point out that the results reported hereafter are robust to slight changes in the parameters in Eqs (IV.4) to (VI.6).

IV.2.3 Epigenetic content of the four prevalent promoter chromatin states

Visualization of the distributions of the thirteen epigenetic marks in each of the four promoter chromatin states in Figs IV.4 and IV.5, shows that most marks are not confined to a single promoter chromatin type. Rather, the four main promoter chromatin types are defined by a unique linear combination of these marks.

P1 (pink): active euchromatin state. More than 90% of the 6 kb promoter loci in P1 are associated (positive enrichment) with histone modifications H3K36me3, H3K4me3, H3K27ac and H3K79me2 (Fig. IV.4), the hallmarks of transcriptionally active euchromatin [6, 55, 166], as well as with RNA polymerase II (Fig. IV.5) and to a slightly less extent with the RPD3-interacting protein Sin3A (Fig. IV.5) as previously found in active euchromatin in *Drosophila* [50]. P1 also regroups the majority of H3K9me1 marked promoter loci consistent with previous observation of higher H3K9me1 levels in the TSS surrounding of active promoters [166]. Most of the promoter regions containing the histone variant H2AZ also belong to P1. The highly conserved histone variant H2AZ has been previously shown to affect nucleosome positioning *in vitro* and *in vivo* [18, 150, 200, 201] and to be associated with chromatin activation *in vivo* [18, 166] by contributing, via nucleosome sliding, to the phasing of a nucleosome free region at TSS [4, 150, 202, 203].

P2 (orange): facultative heterochromatin state. P2 is notably associated with the histone modification H3K27me3 (Fig. IV.4). This mark is well known to be recognized by the chromodomains of Pc proteins and to be implicated in gene silencing [145, 166].

P3 (green): silent "unmarked" heterochromatin. Out of the four promoter chromatin states, P3 corresponds to promoter loci lacking a clear chromatin mark signature. As shown in Figs IV.4 and IV.5, most P3 promoters are not enriched for any available marks. P3 can indeed be compared to the "null" or "black" silent heterochromatin states previously found in *Drosophila* [50, 53] and *Arabidopsis* [51] as covering a significant portion of the genome.

P4 (blue): HP1 associated heterochromatin state. P4 corresponds to the few (679) gene promoters containing the H3K9me3 mark and almost only that repressive mark (Fig. IV.4) as the probable signature of its ability to anchor to the heterochromatin protein HP1 at the origin of establishment of heterochro-

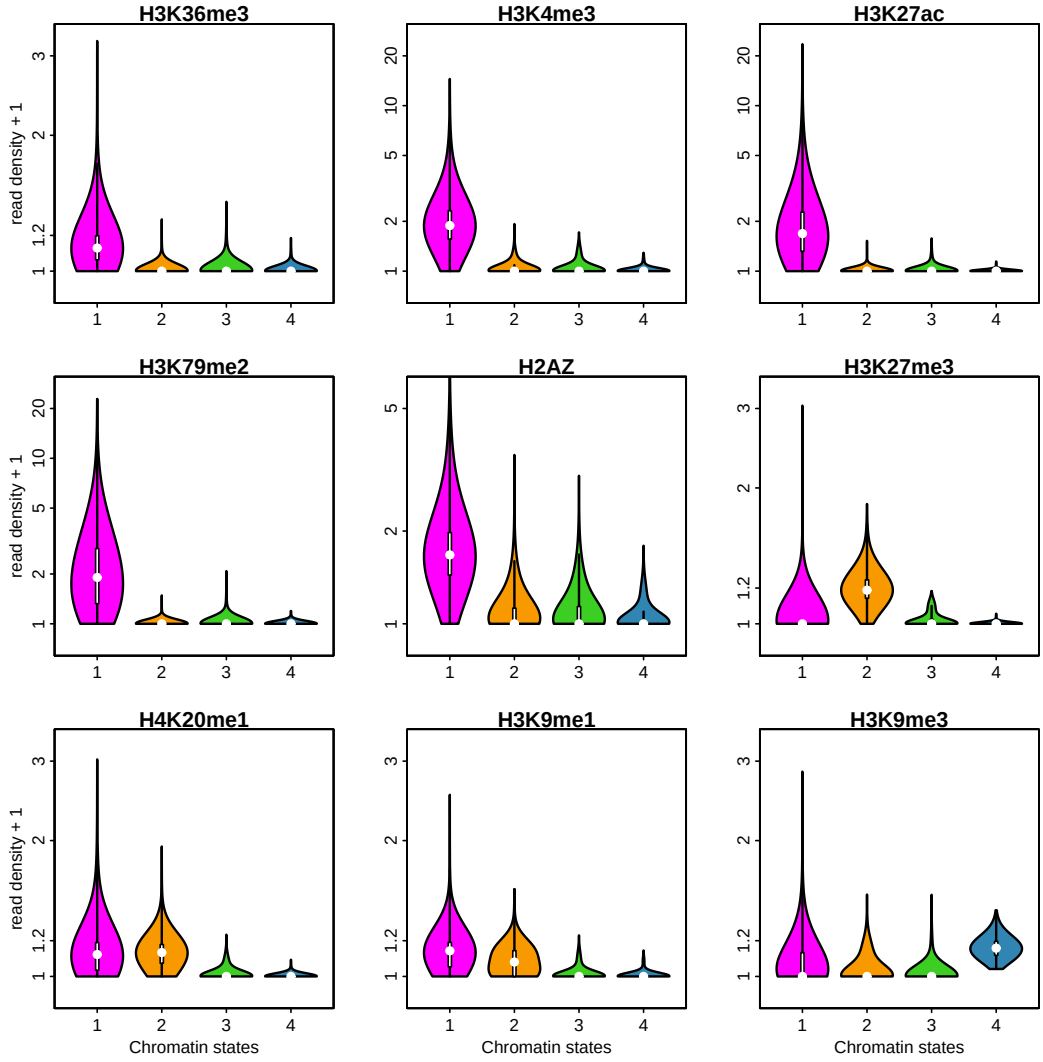

Figure IV.4: Repartition of histone marks in the four promoter chromatin states P1, P2, P3 and P4. Violin plots of the decimal logarithm of histone mark ChiP-Seq read density in 6 kb window around TSS per promoter state. Violin plot combines a boxplot (in white) with a symmetric density plot (colored area). The wider the colored area is the more points are associated with this value. Same color coding as in Fig. IV.3.

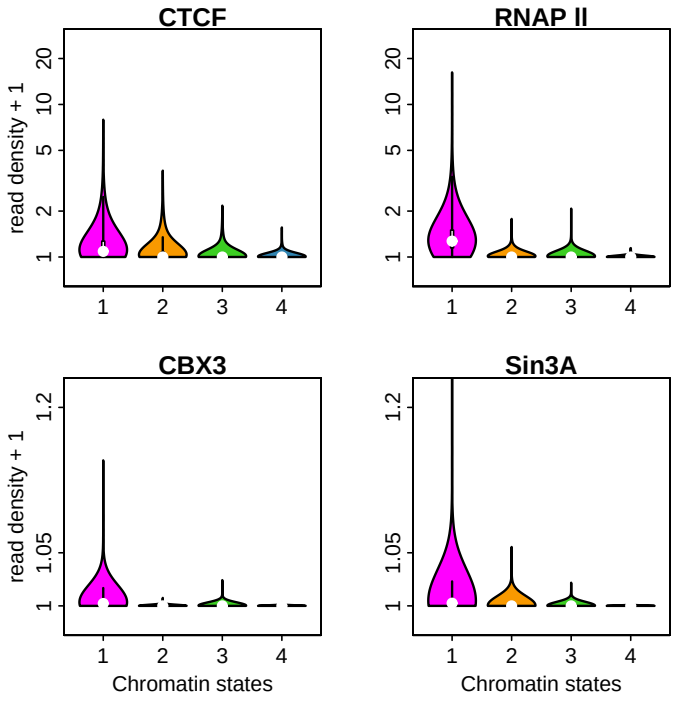

Figure IV.5: Repartition of transcription factors in the four promoter chromatin states P1, P2, P3 and P4. Violin plots of the decimal logarithm of transcription ChiP-Seq read density in 6 kb window around TSS per promoter state. Same color coding as in Fig. IV.3.

matin [145, 166].

Methylation of H3K9 is well known to be implicated in heterochromatin formation and gene silencing [6,55]. The fact that H3K9me1 is found in P1 and to a less extent in P2 and not in P4 (Fig. IV.4) confirms that this epigenetic modification, together with H4K20me1, may also be associated with transcriptional activation [166,195–198]. Note that H3K9me3 is not exclusively found in P4 promoter regions; as seen in Fig. IV.4, 42% of P1 promoters and 25% of the P2 promoters contain some H3K9me3 marks. As mentioned in Sect. III.2.2, when present in combination with all active marks, this mark may drive the anchoring of CBX3 (Fig. IV.5) involved in gene splicing [170–175].

The insulator binding protein CTCF is known to establish chromatin boundaries to prevent the spreading of heterochromatin into transcriptionally active regions [145, 166]. As shown in Fig. IV.5, consistent with this picture, we get, in good agreement with previous observations in *Drosophila* [50, 53], that CTCF is found in P1 promoters and to a slight extent in P2 promoters. This

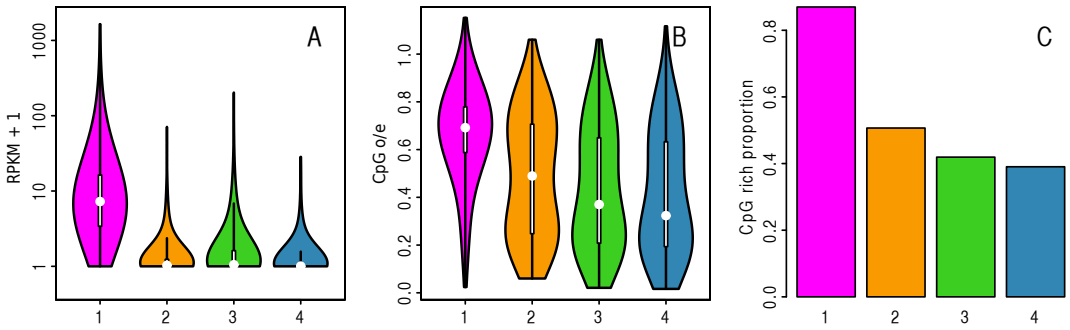

Figure IV.6: Expression level and CpG content in the four promoter chromatin states P1, P2, P3 and P4. (A) Violin plots of the decimal logarithm of RPKM expression score (Eq. IV.1) in the four promoter states. (B) Violin plots of CpG o/e computed in the 6 kb windows around the TSS per promoter states. (C) Proportion of CpG rich genes per promoter state (a promoter is CpGrich if the CpGo/e around its TSS is above 0.48). Same color coding as in Fig. IV.3.

can be understood by that fact that P1 and P2 genes lie together in gene rich, high GC megabase-sized domains of intermingled active euchromatin and facultative heterochromatin regions (see Sect. IV.4).

To summarize, this simple classification into one active promoter chromatin state (P1) and three repressed promoter chromatin state (P2, P3 and P4) of human genes is strikingly similar to those recently reported in *Arabidopsis* [51] and *Drosophila* [50, 53] suggesting the possible existence of some simple principles of epigenetic regulation of eukaryotic genomes.

IV.2.4 A synthetic view of epigenetic regulation of gene activity

Gene expression

As shown in Fig. IV.6A, when investigating gene expression data (Sect. IV.6.1), we find that a vast majority (8312, 88%) of expressed gene promoters with a RPKM > 1 (Eq. (IV.1)) are in the euchromatin state P1. As expected, most (2779, 89%) of the Pc repressed P2 promoters correspond to non expressed genes. Interestingly, we find that the number of non expressed genes in P1 (1250) is non negligible and comparable to the one in P2 (2779). Most of the promoters in the heterochromatin states P3 (3124, 81%) and P4 (609, 91%) correspond to silent genes except a minority of them.

CpG-rich versus CpG-poor promoters

Mammalian promoters can be classified according to their sequence content. Most promoters coincide with regions of high GC content and CpG ratio (or CpG islands) [189, 204–207]. As already noted by others [208–211], the distribution of CpG enrichment is bimodal which is also the case in other mammalian genomes, including the mouse genome. As proposed in a previous work [114], we can use a threshold value r^* (0.48 in Fig. IV.6B,C) so that promoters with a CpG enrichment > 0.48 are considered CpG rich and with CpG enrichment < 0.48 CpG poor (Sect. IV.6.7). These two classes of promoters have different regulations and present different characteristics. Whereas CpG-poor genes have a specific initiation site, usually a TATA-box, CpG-rich genes have a broad initiation site [212]. Besides, CpG-rich promoters evolve more rapidly than CpG-poor ones. A hypothesis on the origin of these two gene categories was proposed in [211] but not investigated further: these two categories could have a different evolutionary history, with CpG-rich genes being the oldest ones, present before the global methylation appeared on vertebrate genomes [204, 207] and CpG-poor being more recent. As shown by the violin plot of CpGo/e in Fig. IV.6B, gene promoter loci in P1 are significantly enriched in CpG as compared to P2, P3 and P4 promoter loci. We clearly find a significant shift of the CpG pdf to smaller values when going from P1 ($\overline{\text{CpG o/e}} = 0.69$)) to P2 ($\overline{\text{CpG o/e}} = 0.49$), P3 ($\overline{\text{CpG o/e}} = 0.37$)) and P4 ($\overline{\text{CpG o/e}} = 0.32$)). Relative to the genome average 0.57, the P1 promoter loci are clearly CpG-rich. In terms of promoter states previously defined, 87% of P1 promoter loci belong to the CpG-rich class as compared to 51% of P2, 42% of P3 and 39% of P4 promoter loci. Thus, a non negligible proportion of gene promoter loci in the repressed heterochromatin states P2, P3 and P4 are CpG-rich but mostly non expressed in K562 human cell line.

IV.3 Interplay between promoter activity and large-scale chromatin environment

IV.3.1 Distribution of promoter states in the four prevalent large-scale chromatin states

In Chapter III [154], we identified four main large-scale chromatin states C1, C2, C3 and C4 that were respectively found in 6572 (23.8%), 5312 (19.2%), 6603 (23.9%) and 6758 (24.4%) loci among the 27656 100 kb loci with a defined MRT. Note that we removed from the analysis 2411 (8.7%) loci that were not properly classified in any of these chromatin states: To address the question of the gene content of these four chromatin states, we used a data set of 17724 genes whose promoters have a valid epigenetic value for the considered 13 epigenetic marks. Some of these genes (1832) were not taken into account in our analysis because their promoters did not belong to any C1, C2, C3 or C4 100kb loci. The mean density of the 15892 genes that belong to one of the four large-scale chromatin states is 6.25 promoters per Mb. As reported in Tables IV.1 and IV.2, the early replicating active euchromatin state C1 is highly enriched in gene promoters (14.82 promoters/Mb) and harbours 69.5% of gene promoters even though it represents about 25% of the total genome coverage by the four large-scale chromatin states. The mid S facultative heterochromatin state C2 also contains a non negligible percentage (17.2%) of gene promoters that indeed corresponds to a modest density 4.74 promoters/Mb. The late replicating unmarked and HP1-associated heterochromatin states C3 and C4 are genuinely gene poor with a very low gene densities 2.34 promoter/Mb and 1.11 promoter/Mb for a total of 8.6% and 4.7% of gene promoters respectively. Let us point out that the mean gene length increases gradually from C1 (42.5kb), to C2 (59.4kb), C3 (83.5kb) and C4 (133.1kb), which explains why the gene coverage decreases less abruptly than the promoter density, with C1 mainly genic (62.9%), C2 modestly genic (49.8%) and C3 (39.5%) and C4 (29.3%) mostly intergenic.

As reported in Table IV.3, when comparing the data in Table IV.2 and the expected promoter number if the probability of belonging to any promoter state Pi were independent from the probability of being in the chromatin state Cj, we find observed/expected ratio values significant greater than 1 for the four (Pi/Ci) associations as the signature of an increasing dependency from

	C1	C2	C3	C4
P1	11.8	0.53	0.1	0.06
P2	1.29	2.98	0.19	0.01
P3	1.6	1.18	2.03	0.29
P4	0.13	0.05	0.02	0.75

Table IV.1: Density of promoters per Mbp in the four large scale chromatin states C1, C2, C3 and C4, for the four epigenetic promoter states, P1, P2, P3 and P4.

	C1	C2	C3	C4
P1	8797	304	57	41
P2	961	1721	113	7
P3	1193	682	1191	191
P4	99	26	14	495

Table IV.2: Number of promoters P1, P2, P3 and P4 in large scale chromatin states C1, C2, C3 and C4.

	C1	C2	C3	C4
P1	1.38	0.19	0.07	0.1
P2	0.49	3.57	0.47	0.05
P3	0.53	1.22	4.23	1.27
P4	0.22	0.24	0.26	16.91

Table IV.3: Observed/expected ratio of a promoter Pi to be in a large-scale chromatin state Cj. The expected number is given by $\frac{n_{Pi} * n_{Cj}}{N}$ where n_{Pi} is the number of promoters in Pi, n_{Cj} the number in Cj and N the total number of promoters.

(P1/C1) (1.38), to (P2/C2) (3.57), (P3/C3) (4.23) and (P4/C4) (16.91). In contrast, the observed/expected ratio values obtained for the $(\text{Pi,Cj})_{i \neq j}$ associations are all smaller than 1 as an indication of some anti-correlation except for (P3, C2) (1.22) and (P3, C4) (1.27) that shows that unmarked P3 promoters are more abundant than expected in both the facultative C2 and C4 heterochromatin states.

	from C1	from C2	from C3	from C4
to P1	0.79	0.11	0.04	0.06
to P2	0.09	0.63	0.08	0.01
to P3	0.11	0.25	0.87	0.26
to P4	0.01	0.01	0.01	0.67

Table IV.4: Transition matrix from large-scale chromatin states to promoter states. Probability of being classified in the promoter state Pi knowing that the promoter is embedded in the large scale chromatin state Cj.

	to C1	to C2	to C3	to C4
from P1	0.96	0.03	0.01	0
from P2	0.35	0.61	0.04	0
from P3	0.37	0.21	0.37	0.05
from P4	0.16	0.04	0.02	0.78

Table IV.5: Transition matrix from promoter states to large scale chromatin states. Probability that a promoter in the class Pi to be embedded in the large scale chromatin state Cj.

IV.3.2 Conditional analysis of promoter activity and large-scale chromatin environment

In Table IV.4, we have expressed the results reported in Table IV.2 in terms of the probability of a promoter to be classified in the promoter state Pi knowing that it is embedded in the large-scale chromatin state Cj. The large scale unmarked C3 and HP1-associated C4 states likely corresponding to nuclear lamina pericentric heterochromatin [145, 166, 213] only contain silent genes with P3 and P4 promoters (∼90%). If large-scale transcriptional activity in C1 euchromatin state is recovered in a large majority (∼80%) of genes with P1 promoters, it does not exclude the presence of inactive genes with P2 (9%) and P3 (11%) promoters. Large-scale facultative heterochromatin state C2 is not very predictive of promoter states since besides a majority of Pc repressed P2 gene promoters (63%) it also contains a significant and non negligible proportion of silent unmarked P3 (25%) and of active P1 (11%) promoters.

Reciprocally, when revisiting the results in Table IV.2 in terms of the prob-

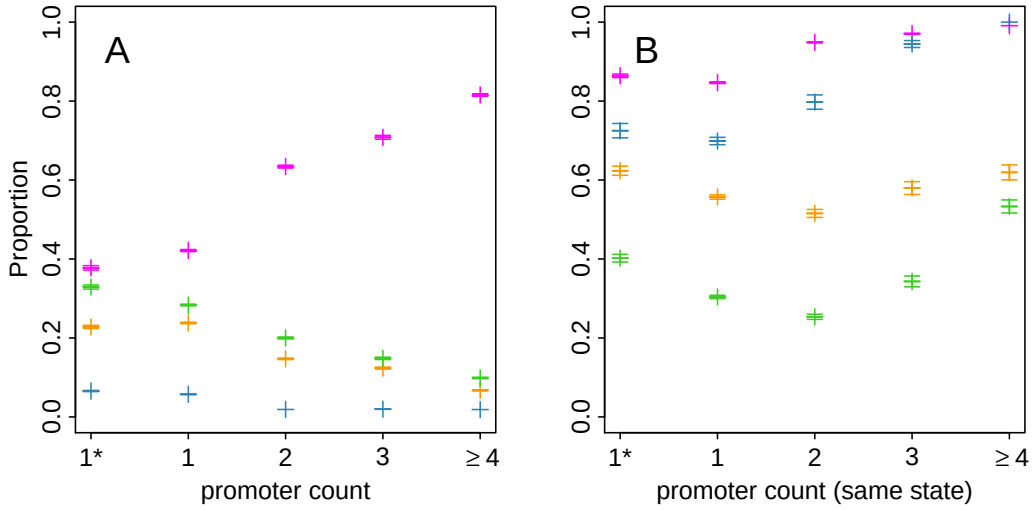

Figure IV.7: Effect of local promoter density on large-scale chromatin state. Promoter count for a gene is the number of promoters that fall in a 100 kb window centered around its TSS. The more the promoter count is high, the more gene rich is the surrounding region. A promoter count of 1* means that the gene is isolated and that its length is smaller than 50 kb (so that the surrounding of this gene is mostly intergenic). Promoter count (same state) is the promoter count taking into account only genes with the same promoter state as the considered gene. (A) Proportions of promoter states P1, P2, P3 and P4 with respect to promoter count. (B) Proportion of promoters with a large chromatin state corresponding to their promoter state (e.g. P1 in C1, etc.) with respect to promoter count (same state). Same color coding as in Fig. IV.3.

ability of a promoter in a given promoter state Pi to be in large-scale chromatin environment Cj, we find in Table IV.5 that with a very high probability (96%) P1 promoters have an active euchromatin C1 environment. This contrasts with the Pc repressed P2 promoters that in majority (61%) belong to the corresponding large-scale facultative heterochromatin C2, but with a significant proportion of them (35%) that are contained in an active C1 environment. The unmarked P3 promoters are rather evenly distributed in C1 (37%), C2 (21%) and C3 (37%). Let us point out that the poorly populated P4 promoter state is consistently found in majority (78%) in the corresponding constitutive heterochromatin state C4 but also in the gene rich euchromatin state C1 (16%) where 1/3 (resp. 2/3) of them are expressed (resp. silent) genes.

Further understanding of these results can be obtained when taking into account gene density. As shown in Fig. IV.7A, when classifying promoters according to gene promoter number in their 100 kb neighbohood, we see that the

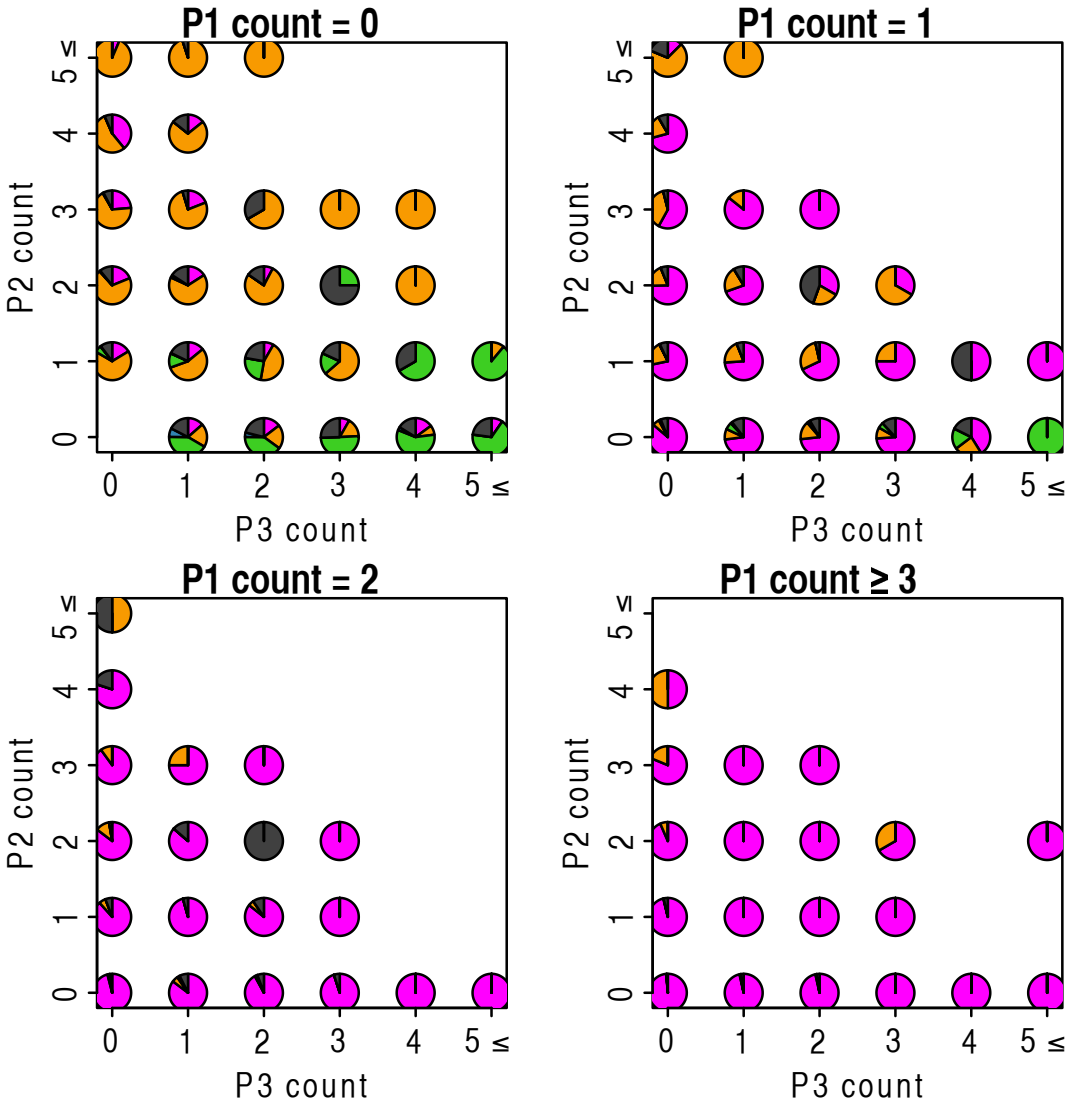

Figure IV.8: Large-scale chromatin states with respect to promoter counts. Pi count of a given gene is the number of Pi promoters that fall in a 100 kb window centered around its TSS. Each panel corresponds to a different active P1 promoter count. For each possible values of the three considered promoter counts (P1, P2, P3), we calculated the proportions of large-scale chromatin states C1 (pink), C2 (orange), C3 (green) and C4 (blue); these proportions are represented by a pie chart. Because P4 promoter state is poorly populated (Table IV.2), we have fixed P4 count=0.

proportion of active P1 promoter increases when increasing the local promoter count to the expense of the proportions of inactive P2, P3 and P4 promoters. Even more spectacular, similar tendancies are observed in Fig. IV.7B when considering now the relative proportions of consistent pairing (Pi, Ci) of a promoter Pi embedded in the corresponding large-scale chromatin environment Ci, when increasing the local density of promoters of the same state Pi. As expected the proportion of transcriptionally active pairing (P1, C1) increases when the 100kb windows surrounding a P1 promoter contains more and more P1 promoters. Naively we would have expected the same increase in the probability of an inactive promoter P2, P3 or P4 to be embedded in the corresponding heterochromatin environment C2, C3 or C4 respectively, when enriching its 100kb neighborhood in promoters belonging to the same promoter state. However, this is only true for HP1-associated promoters. This observation is consistent with P4 promoters being mostly in a separated nuclear compartment (Table IV.5). For promoter states P2 and P3, the pairing (Pi,Ci) doesn't increase with promoter density. Indeed, as shown in Fig. IV.8 (upper left panel), this is only true if this neighborhood contains no P1 promoter. As soon as one or more P1 promoters belong to the neighborhood of a P2 or P3 promoter, then the probability for this promoter to be embedded in the gene rich euchromatin state C1 increases (Fig. IV.8, other panels), which explains the observed behavior of the proportions of inactive pairing (P2, C2) and (P3, C3) in Fig. IV.7B. On the contrary to P1 promoter, the presence of one P4 promoter doesn't imply a C4 environment suggesting that a P4 promoter is not sufficient to drive the association with the pericentric compartment (data not shown). Altogether these results confirm that gene density is a key parameter underlying the coherence between promoter activity and large-scale chromatin environment.

IV.4 Repartition of promoter chromatin states along human chromosomes

IV.4.1 Distribution of promoter chromatin states inside replication timing U-domains

When first concentrating on the gene distribution inside the 876 replication timing U-domains previously identified in K562 cells [94], we reveal a remarkable organization of the four prevalent promoter chromatin states. This is particularly patent in Fig. IV.9A-D where the 876 U-domains were centered and ordered vertically from the smallest (top) to the largest (bottom) and only gene promoters were represented. By simple visual inspection, we recognize in Fig. IV.9A the edges of the U-domains from the local enrichment of active P1 promoters that are mainly confined in a closed ($\sim$150 kb) C1 neighborhood of the "master" replication origins that border these replication domains. Note that this result is quite consistent with the previous observation [114] that CpG-rich gene promoters that are likely to be active in the germ line and do present an important transcription-associated nucleotide compositional asymmetry [105,214–216], also lie preferentially nearby the edges of replication skew N-domains. In Fig. IV.9B, the Pc repressed P2 promoters are mostly found at finite distance ($\sim$200-300 kb) from U-domain borders whose centers are significantly devoided of P2 promoters. In small U-domains ($<$ 1.2 Mb), P2 promoters mainly occupy their centers that are replicated in mid-S phase. In contrast unmarked P3 promoters do not seem to have any preferential positioning inside U-domains where they look like rather homogeneously distributed as shown in Fig. IV.9C. Despite their small number, inactive HP1-associated P4 promoters are mostly found in the central region of large ($>$ 1 Mb) U-domains in Fig. IV.9D; they consistently lie in a late replicating heterochromatin C4 environment. As confirmed on the corresponding mean occupation profiles in Fig. IV.9E, this remarkable organization of gene promoters inside U-domains is consistent with the gradient of chromatin states observed across these replication domains, from C1 at U-domain borders followed by C2, C3 and C4 at centers. Note that as shown in Figs IV.9F and IV.9G, a similar organization is found for CpG-rich and CpG-poor promoters respectively, except that CpG-poor P1 promoters are about one order of magnitude less numerous than CpG-rich P1 promoters.

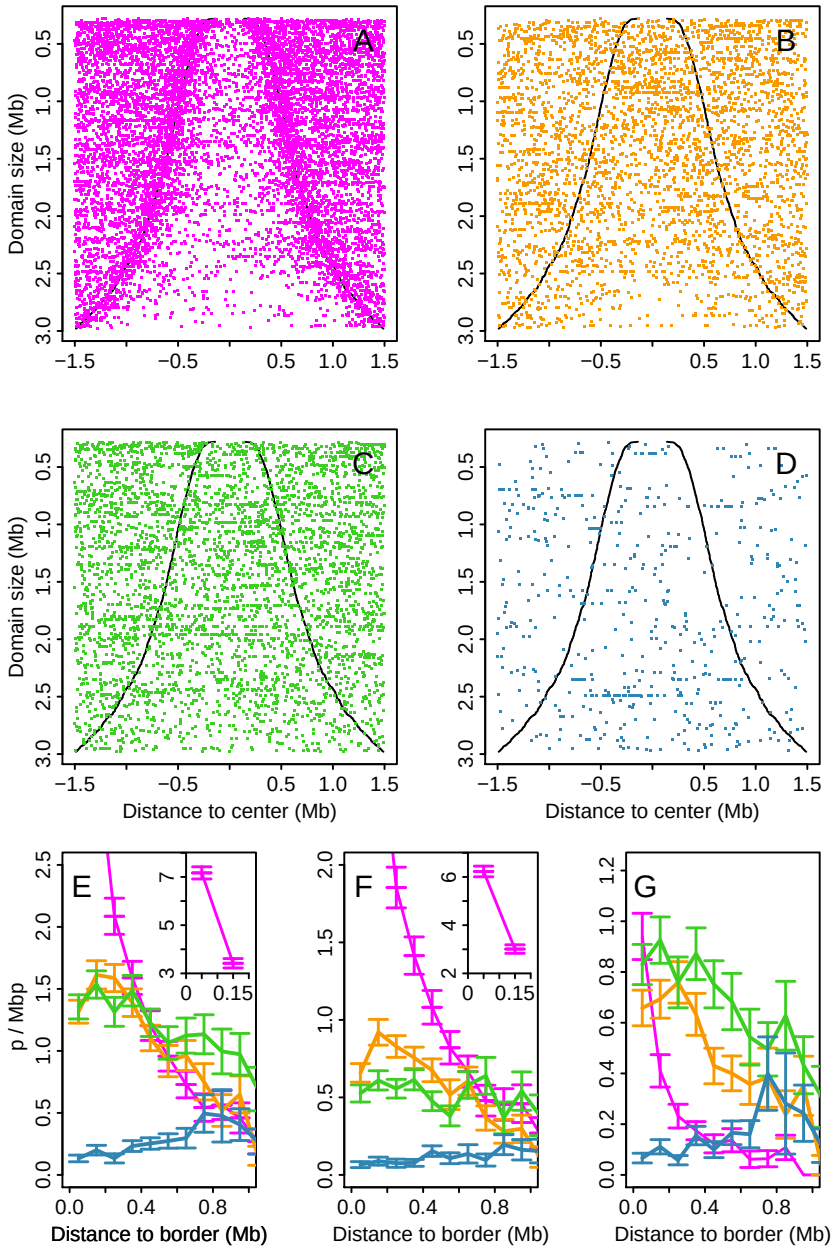

Figure IV.9: Distribution of promoter states inside replication U-domains. (A) The 876 K562 U-domains were centered and ordered vertically from the smallest (top) to the largest (bottom). All active P1 promoters are represented by a dot (pink). (B) Same as (A) for Pc repressed P2 promoters (orange). (C) Same as (A) for the unmarked promoters P3 (green). (D) Same as (A) for HP1-repressed promoters P4. (E) Mean promoter density with respect to the distance to the closest U-domain border. Error bars represent standard deviation. Same color coding as in (A-D). (F) Same as (E) for CpG rich genes. (G) Same as (E) for CpG poor genes.

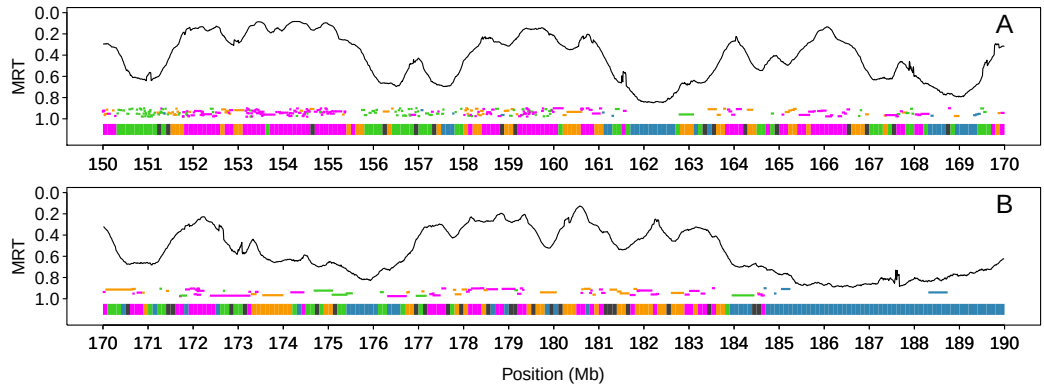

Figure IV.10: Distribution of promoter states along the MRT profile. (A) K562 MRT profile along a 20 Mb long fragment of human chromosome 1. Below the MRT profile, gene positions are indicated by a segment. The segment color indicates the promoter state. Same color coding as in Fig. IV.9. At the bottom of the plot, the chromatin state of each 100 kb window is represented with the following coding: active euchromatin state C1 (pink), Pc repressed facultative heterochromatin C2 (orange), silent unmarked heterochromatin state C3 (green) and HP1-associated heterochromatin state C4 (blue) [154]. (B) Same as (A) for the following 20 Mb fragment of the human chromosome 1.

IV.4.2 Distribution of promoter chromatin states outside replication U-domains

Replication timing U-domains actually cover about 50% of the human genome. In Chapter III [154], we have shown that the other half of the human genome is more in agreement with the dichotomic picture proposed in early studies of the mouse [34, 35, 86] and human [12, 36, 90] genomes, where early and late replicating regions occur in separated compartments of open and close chromatin respectively.

* High GC, gene rich (C1+C2) blocks: About 25% of the human genome (Table IV.6) are covered by megabase-sized GC-rich (C1+C2) chromatin blocks that on average replicate early by multiple almost synchronous origins (*e.g.* the region from 151.5 Mb to 155.8 Mb of human chromosome 1 in Fig. IV.10A)). As reported in Table IV.6, these regions are gene rich with a high density of P1 promoters (6.85 promoters/Mb) and a significant density of P2 promoters (2.15 promoters/Mb) that replicate slightly earlier than the mid-S phase P2 promoters found in replication timing U-domains. Some unmarked P3 promoters (1.41 promoters/Mb) also

	U-domains	C1+C2	C3+C4
total length (Mb)	1293.9	750.6	745.5
mean length (kb)	1431.3	561.8	723.1
promoter number			
P1	3029	6224	197
P2	1550	1449	103
P3	1656	1218	826
P4	306	70	285
density of promoters per Mb			
P1	2.3	8.29	0.26
P2	1.20	1.93	0.14
P3	1.28	1.62	1.1
P4	0.24	0.09	0.38

Table IV.6: Distribution of promoter chromatin states P1, P2, P3 and P4 inside replication U-domains, (C1+C2) blocks and (C3+C4) blocks [154].

belong to these (C1+C2) blocks and correspond to the sub-class of genes with P3 promoters that are expressed in K562. Only few P4 promoters (0.09 promoters/Mb) are found in these early replicating (C1+C2) block regions.

* Low GC, gene poor (C3+C4) blocks: The last 25% of the human genome correspond to megabase-sized GC-poor domains of interspersed (C3+C4) heterochromatin states or of long C4 domains that on average replicate late by again multiple almost coordinated origins (*e.g.* the region from 185 Mb to 190 Mb of human chromosome 1 in Fig. IV.10B). As reported in Table IV.6, these regions are gene deserts with, relatively to their genome mean densities, almost no P1 (0.17 promoters/Mb) and P2 (0.10 promoters/Mb) promoters, and in contrast contain most of the P4 promoters (0.43 promoters/Mb) as well as a significant proportion of P3 promoters.

As reported in Fig. IV.6, P1 and P2 promoters are in large majority CpG rich, which further indicates that C1+C2 blocks are enriched in CpG-rich gene promoters consistent with previous observations that CpG-rich genes tend to

be seated in high GC isochores [154]. In contrast, C3+C4 blocks, as the low GC isochores counterpart, contain only a few genes mostly inactive and with a CpG-poor promoter.

IV.5 Conclusion/Perspectives

In summary, the integrative analysis of epigenetic mark maps in the myelogenous leukemia human cell line K562 has shown that, at the gene promoter scale (±3kb around TSS), the combinatorial complexity of these epigenetic data can be reduced to four prevalent promoter chromatin states that display remarkable similarties with those found in different cell types in *Drosophila* [53] and *Arabidopsis* [51]: P1 regroups all the marks of transcriptionally active chromatin and corresponds to CpG-rich promoters of highly expressed genes; P2 is notably associated with the histone modification H3K27me3 that is the mark of Pc repressed falcultative heterochromatin; P3 corresponds to promoters that are not enriched in any marks as the signature of silent heterochromatin; and P4 characterizes the few gene promoters that contain only the HP1-associated histone modification H3K9me3. When analyzing the coherence between promoter activity (P1, P2, P3 ad P4) and the corresponding large-scale (100kb) chromatin states (C1, C2, C3 and C4) that were shown to replicate at different periods of the S-phase (Sect. III) [154], we confirm gene density as a central parameter underlying the interplay between transcription and replication. Among the striking results obtained about the large-scale chromatin environment from the local knowledge of a gene-promoter activity is the fact that a P1 promoter is almost surely surrounded by an early replicating, gene-rich, transcriptionally active euchromatin state C1. Reciprocally, it is the spreading of the late replicating, gene-poor, HP1-associated heterochromatin large-scale state C4 that almost surely governs the local inactivity of the few unmarked P3 and constitutively silent P4 promoters. When further investigating the spatial distribution of the P1, P2, P3 and P4 promoters along human chromosomes, our study reveals a remarkable gene organization in relation with the MRT. In 50 % of the human genome that are covered by megabase-sized replication U-domains [94, 154], a significant enrichment of highly expressed P1 genes is observed in a closed neighborhood of the early C1 initiation zones that border these domains. P2 promoters are mainly found in the mid-S C2 environment at finite distance (∼200-300kb) from U-domain borders. Inactive P3 and P4

promoters are distributed more homogeneously inside U-domains with a majority of the poorly populated P4 promoter set in the C4 central region of large U-domains likely associated with pericentric nuclear heterochromatin. Thus, in these U-domains where the replication wave starting at bordering "master" replication origins, keeps accelerating thanks to the firing of secondary origins [93], some gradient of gene promoter activity is also observed as the possible consequence of some epigenetic co-regulation of replication and transcription. This intimate relationship between gene activity and MRT is also observed in the other half of the human genome with mainly P1 and P2 promoters in megabase-sized GC-rich and highly genic (C1+C2) chromatin blocks that replicate early in the S-phase, and P3 and P4 promoters in late replicating, gene-poor and GC-poor megabase-sized (C3+C4) blocks.

Extending this study to different cell types including ES, somatic and cancer cells looks very promising. Previous comparative analyses of replication timing profiles during development have revealed important dynamical changes leading to cell type specific patterns of replication [34, 35, 115]. Importantly, these specific replication timing patterns are conserved between human and mouse synthenic regions of related cell types despite the length of evolutionary divergence [36]. Thus MRT profiles likely capture the epigenetic differences between cell types, even when they are closely related, and should be considered as a *bona fide* epigenetic mark [65, 217]. By performing our integrative analysis at low (100 kb) and high (6 kb) resolutions in parallel, we should be in position to investigate the global reorganization of replication domains during differentiation in relation to coordinated changes in chromatin state and gene expression. A number of studies have also demonstrated a clear association between the replication program and cancer genome rearrangement events [218–221]. In particular, MRT was shown to capture important epigenetic modifications involved in genomic misregulation and chromosomal instability during tumoral progression prior to rearrangement events [221]. Extending the present study to cancer cell lines with well defined temporally ordered steps of tumoral progression will provide new knowledge that hopefully will turn out very helpful for cancer diagnosis, prognosis and cancer treatment. This work is under progress.

IV.6 Materials and methods

IV.6.1 Annotation and expression data

Annotation and expression data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC). To construct our data set, we used RefSeq Genes track as human gene coordinates. Genes with alternative splicing were merged into one transcript by taking the union of exons. Hence the TSS was placed at the beginning of the first exon. We obtained a table of 23329 genes. We downloaded expression values from the release 2 of Caltech RNA-Seq track (ENCODE project at UCSC):

`http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeCaltechRnaSeq/`

Expression for one transcript is given in reads per kilobase of exon model per million mapped reads (RPKM) [10]. RPKM is defined as:

$$R = \frac{10^9 C}{NL}, \tag{IV.1}$$

where C is the number of mappable reads that fall into gene exons (union of exons for genes with alternative splicing), N is the total number of mappable reads in the experiment, and L is the total length of the exons in base pairs. We associated 17872 genes with a valid RPKM value in K562.

IV.6.2 Histone marks, H2AZ, CTCF, RNAP II, Sin3A and CBX3 ChIP-Seq data

For all ChIP-Seq data, we downloaded data in the ENCODE standard formats "broadpeaks" and "bigWig" (`http://genome.ucsc.edu/FAQ/FAQformat.html`). Broadpeaks format is a table of significantly enriched genomic intervals. BigWig format is a read count profile at high resolution of 25 bp. Most of the data correspond to the release 3 (August 2012) of the Broad histone track. We downloaded the tables from:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/`

The CBX3 and Sin3A data correspond to the release 3 (September 2012) of the HAIB TFBS track. Tables were downloaded from UCSC:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/`

For the K562 cell line, we downloaded the broadpeak tables for the following antibodies: CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me3, H3K9me3, RNAP ll, H2AZ, H3K79me2, H3K9me1, H4K20me1, CBX3, Sin3A.

IV.6.3 Read density computation around promoters

For each ChIP-Seq data, we filtered the high resolution profiles (BigWig format) by the significantly enriched intervals (Broadpeaks format). Then, for each gene with a valid expression value, the read density was computed as the number of reads that fall in 6 kb window around the TSS divided by the window length. By doing so, we obtained a valid epigenetic value for 13 epigenetic marks around 17724 promoters.

IV.6.4 Rank transformation and Spearman correlation matrix

All statistical computations were performed using the R software (`http://www.r-project.org/`).

In order to compute the Spearman correlation matrix, the read density around promoters was transformed with the R function *rank* with option *ties.method=max*. Then we computed the Pearson correlation matrix on the transformed dataset. To reorder the matrix in Figure IV.1, we computed the Spearman correlation distance $dSCor$ as:

$$dSCor(X,Y) = 1 - SCor(X,Y), \tag{IV.2}$$

where $SCor$ is the Spearman correlation. Then, a dendrogram was computed using the R function *hclust* with option *method=average* and with $dSCor$ as dissimilarity.

IV.6.5 Principal component analysis

Principal component analysis was performed on the rank transformed dataset using the function *dudi.pca* from the R package *ade4* (see `http://pbil.univ-lyon1.fr/ADE-4` and [167]) with the option *scale=TRUE* (*i.e.* each variable is centered and normalized before the PCA computation). The first three components were retained which accounts for 74% of the dataset variance (see Figure IV.2(B,C)), and promoter states were defined in this 3D space.

IV.6.6 Definition of promoter chromatin states

Promoter chromatin states were defined as subdivisions of the 3D principal component space (Figure IV.3). Geometrical definitions of those subdivisions are given below:

$$P_4 = \left\{(x,y,z) \in \mathbb{R}^3 : x > 1.9, y > 0.5, z > 0.9\right\} \tag{IV.3}$$

$$P_3 = \left\{(x,y,z) \in \mathbb{R}^3 : (x-2.6)^2 + (y-1)^2 < 1.4, (x,y,z) \notin P_4\right\} \tag{IV.4}$$

$$P_2 = \left\{(x,y,z) \in \mathbb{R}^3 : y < \frac{4}{3}(x-2), (x,y,z) \notin P_3 \cup P_4\right\} \tag{IV.5}$$

$$P_1 = \left\{(x,y,z) \in \mathbb{R}^3 : y > \frac{4}{3}(x-2), (x,y,z) \notin P_3 \cup P_4\right\} \tag{IV.6}$$

where x, y, z are the values along the first PC1, the second PC2 and the third PC3 principal components, respectively.

IV.6.7 CpG o/e computation and GC content

CpG observed/expected ratio (CpG o/e) was computed as $\frac{n_{CpG}}{L-l} \times \frac{L^2}{n_C n_G}$, where n_C, n_G and n_{CpG} are the numbers of C, G and dinucleotides CG, respectively, counted along the sequence, L is the number of nonmasked nucleotides and l is the number of masked nucleotide gaps plus one, *i.e.* L-l is the number of dinucleotide sites. The CpG o/e was computed over the sequence after masking annotated CGIs.

IV.6.8 100 kb resolution chromatin states

Chromatin states for the myeloid cell line K562 were retrieved from Chapter III [154]. Large scale chromatin states define an epigenetic segmentation

of the human genome in four prevalent chromatin states C1, C2, C3 and C4 respectively, for 27656 100 kb non-overlapping windows. The large scale chromatin state for a gene is the state of the 100 kb window its TSS is embedded in.

IV.6.9 Promoter count definition

Promoter count for a gene is the number of promoters that fall in a 100 kb window centered around its TSS. For each gene we compute five kinds of promoter count for:

* all genes. This give an indication of the gene density around that gene;
* genes which belong to a promoter class giving 4 promoter counts.

IV.6.10 Mean replication timing data and replication U-domain coordinates

Timing profiles for the immature myeloid cell line K562 were obtained from the authors [94]. The mean replication timing (MRT) is given for 27656 100 kb non-overlapping windows in hg18 coordinates. We also retrieved the coordinates of the 876 U-domains in K562 from the authors [94].

Chapter V

Embryonic stem cell specific master replication origins at the heart of the loss of pluripotency

In this Chapter, we extend the integrative analysis made in Chapter III to several human cell lines. In the somatic cell lines, we recover the four prevalent chromatin states found in the K562 cell line. Interestingly, the embryonic stem cell (ESC) line has a singular epigenetic landscape that involves four specific chromatin states: an active gene rich early replicating euchromatin state (EC1), a mid-S accessible chromatin state (EC2) enriched in bivalent genes, a "null" chromatin state (EC3) devoid of epigenetic marks and a gene-poor highly dynamic chromatin state (EC4) hindering heterochromatin compaction both replicating late. Comparative analysis of U-shaped MRT domains in seven cell lines reveals a widespread plasticity of the replication program during differentiation. Besides some epigenetically regulation, master replication origins at MRT U-domains borders that are shared by all cell types are specified by a local enrichment in nucleosome free regions (NFRs) encoded in the DNA sequence suggesting that they have been selected during evolution. The initiation zones specific to the ESC line bear a particular epigenetic signature. Almost equally distributed in the EC1, EC2 and EC4 chromatin states, these early initiation zones (∼200kb) are significantly enriched in the insulator binding protein CTCF and in pluripotency transcription factors (*e.g.* NANOG and OCT4). Surprisingly, the ones in EC4 appear in GC-low, gene desert regions that are locally enriched in the histone variant H2AZ and also in pluripotency factors

NANOG and OCT4. This demonstrates the importance of these epigenetic marks in the regulation of ESC replication independently from transcription. We also emphasize the important role of H2AZ and H3K4me1 in ESCs for maintaining the chromatin in a highly dynamic and accessible state that is refractory to polycomb and HP1 binding. These results shed a new light on the epigenetically regulated global chromatin reorganization that underlies the loss of pluripotency and lineage commitment. Results reported in this chapter were submitted to Nucleic Acids Research [222].

V.1 Introduction

One of the most remarkable phenomenon in biology is the generation of a whole organism containing a large and phenotypically diverse collection of cells and tissues from a single totipotent cell. This tremendous level of diversity in cellular functions originates from a unique genomic DNA sequence. Since the original sequencing of the human genome a decade ago [1], it has become clear that the functional role of the primary DNA sequence is not only to code for proteins which represent less than 5% of the mammalian genomes, but also to contribute to controlling the spatial structure of DNA in chromatin and in turn to regulate nuclear functions including transcription and replication [2,4]. But as development goes on, the use of the DNA sequence has to be altered to enable lineage commitment. Epigenetic mechanisms including DNA methylation [189, 190, 223–227], histone modifications [6–8, 15, 150, 162, 164, 166, 228] and chromatin structure and dynamics [41,54,55,98,155,229–235] have been proposed to play a key role in regulation of embryonic development, the maintenance of pluripotency and self-renewal of ESCs, lineage specification and the maintenance of cellular identity during differentiation [236–240]. For years, transcriptional and chromatin changes during mammalian development have been attracting increasing interest. Among noteworthy advances, let us mention the identification of pluripotency markers including NANOG/SOX2/OCT4 [7, 241] and of trithorax proteins and polycomb complexes [242–247] as major actors in developmental gene regulation, the identification of the neural restrictive silencer factor NRSF that represses transcription of several neuronal genes in neural development [248]. Also, as differentiation progresses, chromatin structure switches from a highly dynamic, accessible and permissive euchromatin in ESCs to a less open chromatin riddled with accumulating highly condensed

transcriptionally inactive heterochromatin regions [236–238, 249–251].

In contrast to this overwhelming activity concerning the interplay between chromatin structure and transcription regulation during development, only little attention has been paid to replication and its potential role in lineage commitment and fidelity. In pioneering studies, in mouse [34, 35] and human [11, 12], replication domains along chromosomes were delineated in constant timing regions (CTRs) of coordinated origin firings and timing transition regions (TTRs) as origin-less regions [252–254]. In good agreement with previous studies in *Drosophila* [50, 88], these CTRs regions in different mammalian cell types revealed a correlation with epigenetic modifications [89]. Early CTRs tend to be enriched in open chromatin marks, whereas late CTRs are mostly associated with repressive HP1-associated marks [34, 36]. Actually, each cell type presents specific replication timing patterns with mouse ESCs showing a clear MRT pluripotency fingerprint [115]. Differentiation induces important changes in MRT profiles in chromosomal units of size $\sim$ 400-800 kb [34–36]. Early to late (EtoL) MRT changes were associated with loss of pluripotency and largely preceded, in development, late to early (L to E) changes associated with germ-layer specific transcriptional activation [36]. Importantly, these dynamic changes in MRT come along with some sub-nuclear repositioning [34–40]. EtoL (resp. LtoE) transitions occur simultaneously with a movement from (resp. towards) interior of the nuclei towards (resp. from) a more peripheral location or near nucleoli [40–44]. Recent experimental studies of long-range chromatin interactions using chromosome conformation capture techniques [26, 30, 36, 90] have confirmed that 3D chromatin tertiary structure plays an important role in regulating replication timing.

However, in contrast with the above dichotomic picture with early and late replicating loci occurring in separated compartments of open and closed chromatin respectively [26, 36, 90], about half of the genome is paved by the replication U-domains where the MRT is U-shaped [94, 101]. Replication U-domains are likely central to genome regulation since the dynamical changes in MRT profiles observed during differentiation [34–36, 115] mainly occur in the 50% of the genome that are covered by U-domains [94]. Indeed the results reported in [94] show that the so-called replication domain "consolidation" phenomenon [34] actually corresponds to the disappearance (EtoL transition) or appearance (LtoE transition) of a U-domain border during differentiation. Overall, these results point out the "master" replication origins at

U-domain borders as a possible clue to the understanding of the plasticity of the spatio-temporal replication program, gene expression and chromatin organization across different cell lines during development, lineage commitment and fidelity.

Here we extend the analysis presented in Chapter III to different cell types including the ESC H1hesc, three hematopoietic cell lines (K562, Gm1278, Monocyte CD14+), a mammary epithelial cell line (Hmec) and an adult fibroblast cell line (Nhdfad). By investigating the global reorganization of replication U-domains in these different cell types in relation to coordinated changes in chromatin state and gene expression, we shed a new light on the chromatin-mediated epigenetic regulation of transcription and replication during human differentiation. Because they are likely to be the cornerstone to better understanding of pluripotency maintenance, developmental specification and lineage fidelity, we will pay special attention to the "master" replication initiation zones that border U-domains and specially to those that are specific to ESCs.

V.2 Results

V.2.1 Combinatorial analysis of chromatin marks

We investigated relationships between the genome-wide distribution of eight histone modifications, one histone variant and one binding protein in five sommatic cell types including an immature myeloid cell line (K562), a monocyte cell line (Monocd14ro1746), a lymphoblastoid cell line (Gm12878), a mammary epithelial cell line (Hmec), an adult dermal fibroblast cell line (Nhdfad) and an ESC line (H1hesc) for which we also considered the ATP-dependent helicase CHD1, the subunit EZH2 of PRC2 (polycomb repressive complex 2) and the two pluripotency transcription factors NANOG and OCT4. As a first step, we computed the Spearman correlation coefficient of each mark with each other (Sect. V.5.6). We next represented the resulting matrix as a heatmap after having reorganized rows and columns with a hierarchical clustering based on the Spearman distance (Eq. V.1) (Figs V.1 and V.2). This analysis was very enlightening since, on the one hand it revealed that the correlation matrices obtained for the five sommatic cell lines strongly ressemble to the one obtained in K562 in our Chapter III [154] (Fig. V.1 and V.2), and on the other hand it clearly discriminated the pluripotent H1hesc cell line for having a drastically

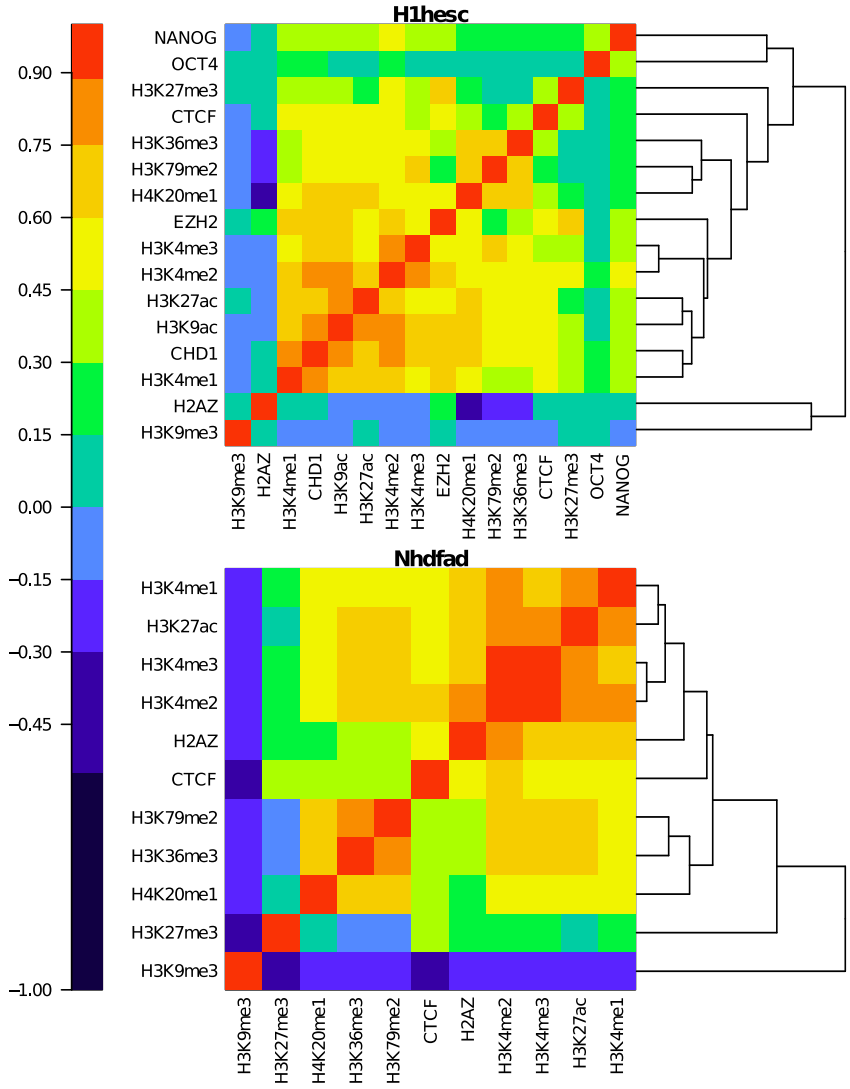

Figure V.1: Spearman correlation matrix between epigenetic marks in H1hesc (top) and Nhdfad (bottom). For each cell line, we computed the Spearman correlation over all 100 kb non-overlapping windows with a valid score. Spearman correlation value is color coded using the color map shown on the left. Lines for the epigenetic marks were reorganized by a hierarchical ordering using Spearman correlation distances (Eq. (V.1)) as illustrated by the dendrograms on the right of the corresponding matrices. This ordering implies that highly correlated epigenetic marks are close to each other.

different correlation structure between epigenetic marks (Fig. V.1).

In the epigenetic mark matrices obtained for the differentiated cell lines Nhdfad (Fig. V.1), Hmec, Monocd14ro1746, K562 and Gm12878 (Fig. V.2), all

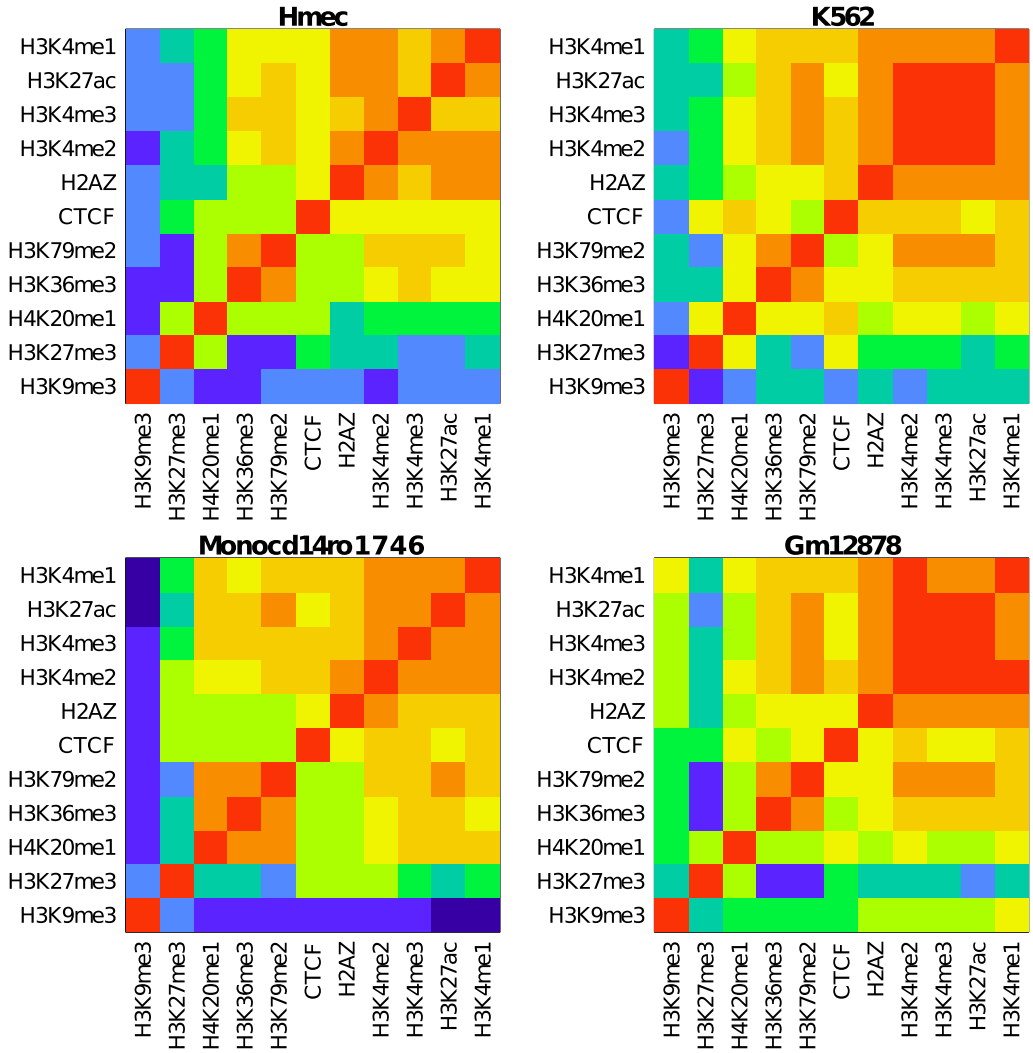

Figure V.2: Spearman correlation matrix between epigenetic marks in Hmec, Monocd14ro1746, K562 and Gm12878 cell lines. Same color map as in Fig. V.1.

histone modifications that are known to be involved in transcription positive regulation, namely H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K36me3, H3k79me2 and H4K20me1, form a block that also includes the histone variant H2AZ and the transcription factor CTCF, meaning that all these marks are all correlated with each other and are likely to occupy similar regions in the genome [164, 166]. In fact, two lines are clearly apart in all correlation matrices as illustrated on the hierarchical clustering dendrogram (Fig. V.1). They correspond to the repressive chromatin marks H3K27me3 and H3K9me3 that are respectively associated with the so-called facultative and constitutive heterochromatins [154]. These two marks are recognized by the chromodomains of polycomb (Pc) proteins and heterochromatin protein 1 (HP1) respectively, components of distinct gene silencing mechanism which may explain that they are anti-correlated with each other. While H3K9me3 behaves quite independently if not anticorrelated with most of the active chromatin marks (except for Gm12878 where some positive correlations were observed), H3K27me3 correlates to some of them in a cell line dependent fashion but quite systematically to CTCF (Figs V.1 and V.2). This consistency of epigenetic mark correlations in the five differentiated cell lines prompted us to build a "shared" epigenetic space (Sect. V.5.5). This consisted in pooling data points of all differentiated cell lines together and then in applying the PCA and clustering algorithm to reduce the dimensionality of the data. As previously experienced with K562 in Chapter III [154], we concentrated on the first four principal components which together account for 86% of the total data set variance (Fig. V.3A'). By projecting the 100kb genomic loci on the (PC1, PC2) plane (Fig. V.3C') and (PC3, PC2) plane (Fig. V.3B'), we noticed that four areas contain most of the population. On the (PC1, PC2) plane a large area of medium density comes out from a plane of much higher density where, as viewed on the (PC3, PC2) plane, loci roughly lie along two differently populated straight lines with a very high density of loci concentrated at the intersection of these lines. This led us to fix the number of clusters to four in the Clara algorithm [153] (Sect. V.5.8), as confirmed when using the within-cluster sum and gap statistical criteria (Sect. III.4.6) [154]. When labeling each of the four main chromatin states with a color, we obtained four domains in the (PC1, PC2, PC3, PC4) space that have common boundaries as illustrated on the (PC1, PC2) projection plane (Fig. V.3D',E'). To improve the quality of our clustering procedure, we filtered out poorly clustered data points that were closer to another cluster than the one they belong to and had a negative silhouette [168] (Sect. III.4.6).

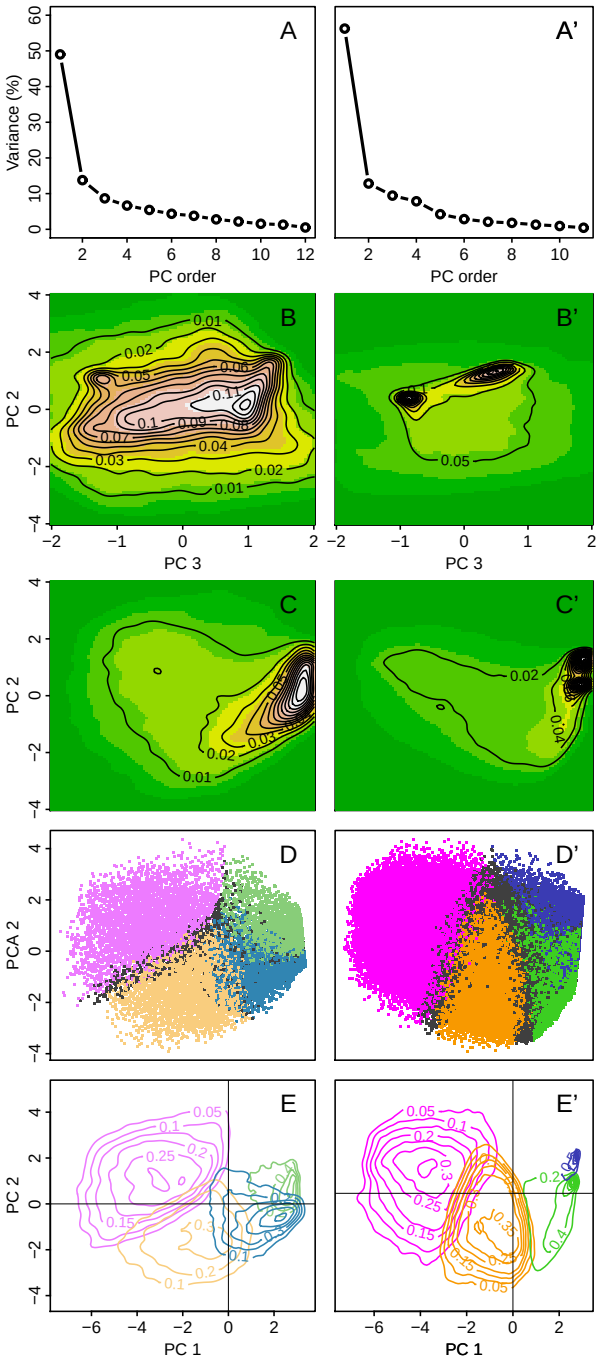

Figure V.3: PCA analysis and clustering procedure for ESC line (A-E) and the five differentiated cell lines (A'-E'). (A, A') Percentage of variance accounted by the eleven principal components ordered according to their corresponding variance (eigenvalues). (B, B') Two-dimensional (2D) density of the data on the plane defined by the second (PC2) and third (PC3) principal components. The densities were computed by a kernel density estimation. (C,C') Projection of the data on the (PC1, PC2) plane. (D, D') Scatterplot of the data points when projected on the (PC1, PC2) plane; color dots indicate the four chromatin states as found by our clustering procedure. (E, E') Density of data points on the (PC1, PC2) plane using the same color coding as in (D, D'). In (D, E) the colors have the following meaning: EC1 (light pink) transcriptionally active chromatin, EC2 (light orange) bivalent chromatin, EC3 (light green) silent unmarked chromatin, EC4 (light blue) dynamically accessible chromatin poised to HP1-heterochromatin expansion. In (D', E') the colors correspond to: C1 (pink) transcriptionally active chromatin, C2 (orange) chromatin repressed by polycomb, C3 (green) silent unmarked chromatin, C4 (blue) HP1 heterochromatin. In (D, D', E, E') the points in dark grey are not classified in any chromatin state (Sect. V.5.8).

The epigenetic mark correlation matrix obtained for the pluripotent H1hesc cell line (Fig. V.1) displays drastic differences from the ones previously obtained for differentiated cell lines. Among others, let us mention the repressive polycomb-associated mark H3K27me3 which now strongly correlates with most of the active marks and specially H3K4me3 as the probable signature of bivalent ESC chromatin [166, 228, 236, 238, 242, 243]. Also the histone variant H2AZ that now correlates as much with both the repressive marks H3K27me3 and H3K9me3 as with some of the active marks, which likely is an indication of specific highly dynamic and accessible chromatin of pluripotent cells [6,166,228,236,238,251]. When reproducing our PCA and clustering analysis on the H1hesc epigenetic data, we again found that four PCs were enough to account for 86% of the total variance (Fig. V.3A), and that one could still reduce the ESC epigenetic complexity to four chromatin states (Fig. V.3B-E) but, as described in the next sub-section, these chromatin states are distinct from the ones delineated in sommatic cells confirming that ESCs and differentiated cells have different epigenomes [6, 15, 166, 228, 236, 251].

V.2.2 Epigenetic content of prevalent chromatin states in ESCs versus differentiated cells

The four chromatin states so identified in the five differentiated cell lines are quite similar to the ones previously found in K562 in chapter III [154] (see also [238]). C1 is a transcriptionally active chromatin state enriched in the histone modifications H3K27ac, H3K4me1, H3K4me3, H3K36me3 (Fig. V.4) and H3K4me2, H3K27me2, H4K20me1 (Fig. V.5), as well as in the histone variant H2AZ whose binding level was shown to correlate with gene activity in human [166] (Fig. V.4). C2 is notably associated with the histone modification H3K27me3 (Fig. V.4) and hence corresponds to a polycomb repressed chromatin state [145,166]. C3 can be compared to the "null" or "black" silent heterochromatin regions devoid of chromatin marks previously found in *Drosophila* [50,53] and *Arabidopsis* [51]. C4 corresponds to the HP1-associated heterochromatin state with all C4 100 kb-loci containing H3K9me3 and almost only that repressive mark (Fig. V.4) [145, 166]. Note that CTCF which is known to establish chromatin boundaries to prevent the spreading of heterochromatin into transcriptionally active regions [142, 176] was found in C1 and to a slightly less extend in C2 loci (Fig. V.4).

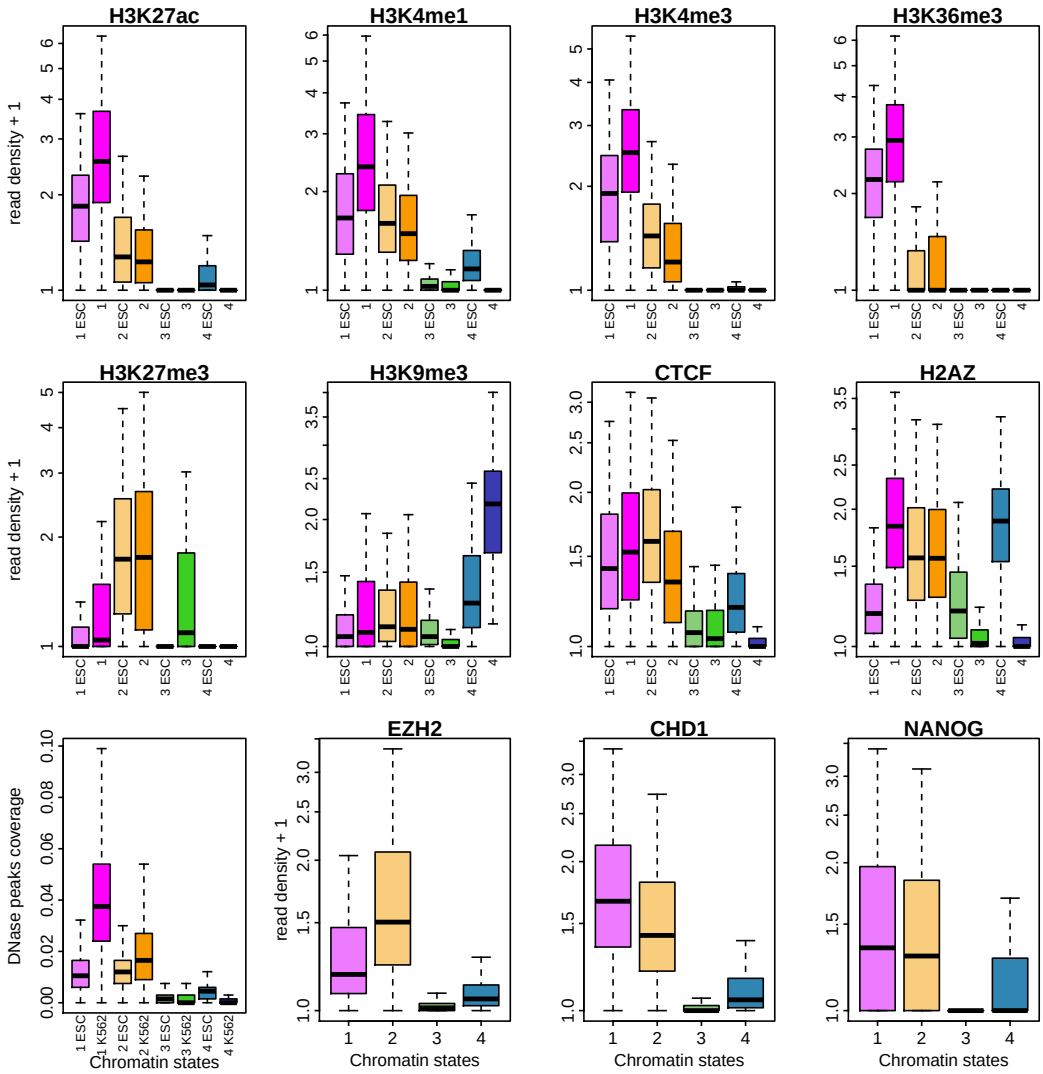

Figure V.4: (First two rows) Repartition of epigenetic marks in the four prevalent chromatin states of H1hesc cell line (EC1, EC2, EC3, EC4, same color coding as in Fig. 1D and E) and differentiated cell lines (C1, C2, C3, C4, same color coding as in Fig. 1D' and E'). Boxplots of the decimal logarithm of histone mark ChiP-seq read density in 100 kb non-overlapping windows per chromatin state. (Third row) Boxplots of the coverage of DNase1 hypersentive peaks in 100 kb non-overlapping windows per chromatin state in H1hesc and K562 cell lines and the decimal logarithm of EZH2, CHD1 and NANOG ChiP-seq read density in 100 kb non-overlapping windows per chromatin state in h1Hesc cell lines. Same color coding than above.

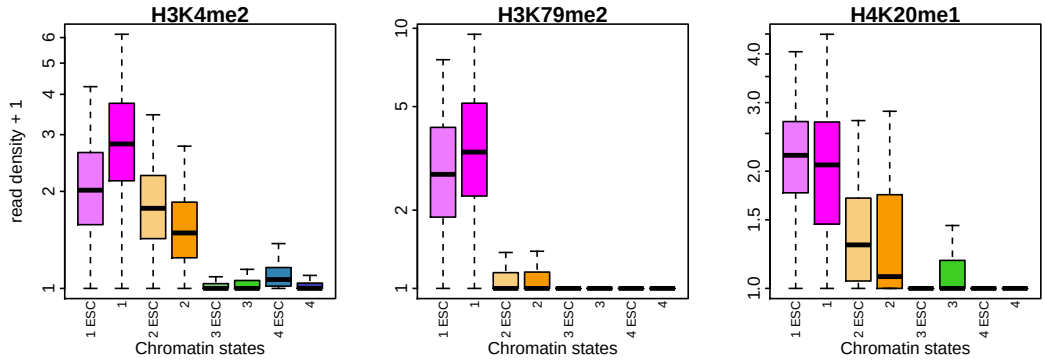

Figure V.5: Repartition of the histone modifications H3K4me2, H3K79me2 and H4K20me1 in the four prevalent chromatin states of H1hesc cell line (EC1, EC2, EC3, EC4, same color coding as in Fig. 1D and E) and differentiated cell lines (C1, C2, C3, C4, same color coding as in Fig. 1D' and E'). Boxplots of the decimal logarithm of epigenetic mark CHip-seq read density in 100 kb non-overlapping windows per chromatin state.

Chromatin states in pluripotent H1hesc cell line (EC1, EC2, EC3, EC4) are different even though they display some similarities with the above described differentiated chromatin states (C1, C2, C3, C4). As for differentiated C1 state but to a slightly less extent, more than 75% of 100kb-loci in EC1 state contain all the active histone modification marks considered (Figs V.4 and V.5). More than 75% of 100 kb loci in EC2 like in C2 are marked by H3K27me3 which is deposited by polycomb complex PRC2 and then enhances PRC1 targeting [245, 247,255] (Fig. V.4). Consistently, EZH2, which is a subunit of PRC2 containing a SET domain that acts upon H3K27 as a methyltransferase, was abundantly found in EC2 confirming the polycomb activity of this state. CTCF is also present in both EC1 and EC2 as previously seen in C1 and C2 but in slightly reverse importance, EC2 being more enriched than C2 and vice versa for EC1 and C1 (Fig. V.4). C1, C2 and EC1, EC2 being the most genic chromatin states in differentiated and ESCs, this result is coherent with the correlation observed between CTCF positioning and gene density [256]. H4K20me1 which was recently shown to strongly correlate with gene activation [166], was consistently found in EC1 and C1 but more surprisingly also in EC2 and C2 which are silent chromatin states (Fig. V.5). Interstingly, recent worls have confirmed that PR-Set 7 involved in the deposition of H4K20me1 plays an important role in the control of replication origin firing in mammalian cells [178–180].

However the epigenetic chromatin states in pluripotent ESCs and differentiated cells bear more differences than similarities. Systematically the differen-

tiated C1 state is more enriched in active histone marks than the pluripotent EC1 state, and this for all histone modifications but H4K20me1 (Figs V.4 and V.5). Relatively to EC1, EC2 contains more H3K4me3 than C2 relatively to C1 (Fig. V.4), which, with the enrichment of EC2 in H3K27me3, is an indication of bivalent heterochromatin. But the most striking difference concerns the pluripotent state EC4 whose epigenetic content is qualitatively and quantitatively different from the one of C4. Noticeably, H2AZ is highly present in more than 75% of EC4 100kb loci which contrasts with its scarity in C4 (Fig. V.4). As compared to C4 which is enriched in the HP1-associated heterochromatin mark H3K9me3, EC4 contains significantly less H3K9me3 as a possible compensation of the enrichment in H2AZ (Fig. V.4). As recently observed in human [238], the enrichment of the ESCs in the histone variant H2AZ associated with nucleosome exchange and remodeling [166,202,235,257] is likely to contribute to the highly dynamic properties of pluripotent chromatin and its refractory character to both HP1- and polycomb heterochromatin restriction [6,166,236,238,249]. This interpretation is stengthened by the observation that in contrast to C4, EC4 is enriched in CTCF (Fig. V.4), which besides its insulator properties [142, 176], is also known to mediate long-range intra- an inter- chromosomal interactions [176, 256, 258–261]. Thus, the accessible and more relax EC4 chromatin might be more central in the nucleus than the HP1-associated heterochromatin C4 state that likely corresponds to the emergence of compact chromatin at the nuclear periphery [41, 54, 55, 155, 229–236, 251].

To get a better comprehension of ESC chromatin states, we looked at two additional epigenetic marks known for their implication in pluripotency. Globally all chromatin remodelers are over expressed in ESC [262] but only some knockdown are known to impair pluripotency. The ATP-dependent helicase CHD1 is one of these [263]. As reported in mouse [263, 264], CHD1 helps at maintening a globally more loose chromatin in ESCs. Interestingly, CHD1 is present in EC1 and EC2 (Fig. V.4) which makes sense since both these chromatin states contain most of the human genes whose expression can possibly be altered by CHD1 in pluripotent cells [263]. But CHD1 is also present in 75% of EC4 100 kb loci confirming that this remodeler contributes to prevent HP1-associated constitutive C4 heterochromatin spreading and compaction [263]. The pluripotent OCT4/SOX2/NANOG network enables self-renewal properties of ESCs, and ectopic expression of these factors together with additional factors or mechanismes was shown to reprogram somatic cells into pluripotent cells (iPS cells) [265–267]. NANOG was found to the same extend in

Chromatin states	EC1	EC2	EC3	EC4	ED
H1hesc	0.21	0.19	0.27	0.24	0.09
Chromatin states	C1	C2	C3	C4	D
K562	0.21	0.14	0.28	0.26	0.11
Monocd14ro1746	0.23	0.18	0.13	0.30	0.16
Gm12878	0.21	0.15	0.36	0.12	0.16
Hmec	0.22	0.21	0.25	0.16	0.16
Nhdfad	0.21	0.22	0.19	0.25	0.13

Table V.1: Percentages of 28465 sequenced 100-kb windows that belong to the EC1, EC2, EC3 and EC4 chromatin states in H1hesc and to the C1, C2, C3 and C4 chromatin states in differentiated cell lines. ED and D correspond to 100-kb windows that were not classified in any chromatin states (Sect. V.5.8).

EC1 and EC2 (Fig. V.4) which is consistent with the fact that NANOG regulates roughly the same number of expressed genes and silent genes [241, 268]. NANOG is surprisingly present in the gene-poor EC4 state suggesting that it may play a role in promoting the relative openess of this pluripotent chromatin state.

V.2.3 Chromatin state coverages and chromatin state changes between cell lines

When comparing the genome coverages, *i.e* the percentages of the 28465 100 kb non-overlapping windows corresponding to the sequenced part of the human genome that belong to the previously identified prevalent chromatin states, we found that whatever the considered cell line, less than 20% of these windows were not properly classified in any chromatin state (Table V.1). In H1hesc cells, EC1 and EC2 coverages are about the same ($\sim$ 19-21%) and are quite similar to the C1 and C2 coverages ($\sim$ 15-23%) generally observed in the five differentiated cells. If the EC3 (27%) and EC4 (24%) coverages are comparable in the ESCs, the C3 and C4 coverages in the differentiated cells are much more variable from 11% to 36%. The lower C4 coverage found in Hmec (16%) than in Nhdfad (25%) may reflect some difference in the culture environment. As reported in [238], the prevalence of H3K9me3 can be tuned up by growth conditions especially in adherent cultures in the presence of serum or other potent growth stimuli. In contrast to Hmec whose growth medium did not

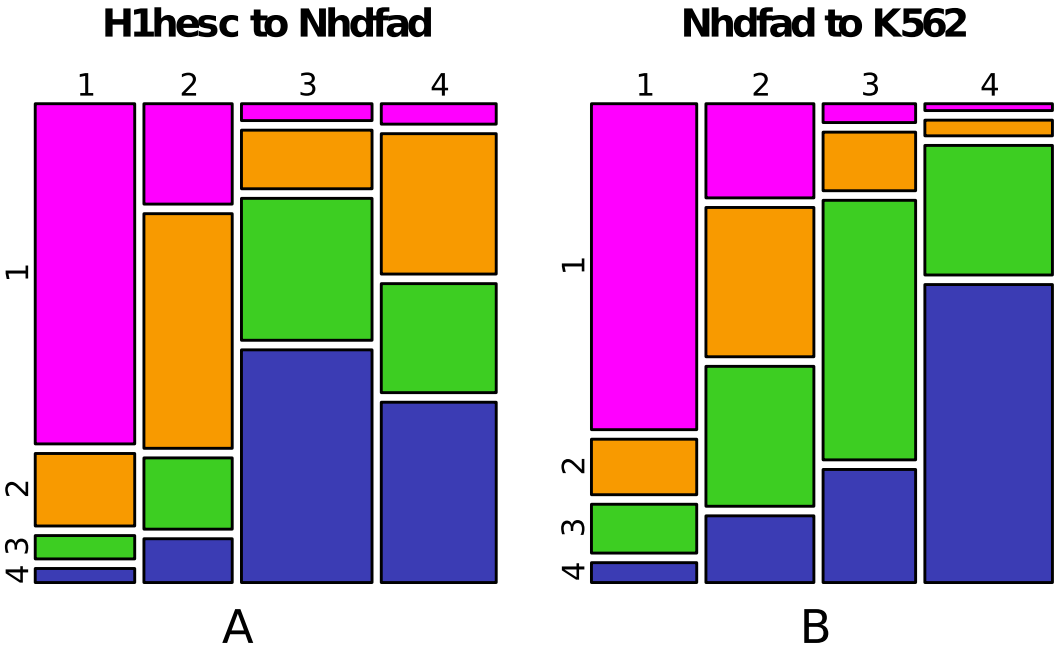

Figure V.6: (A) Mosaic plot representing the probabilities of transition from H1hesc chromatin states to Nhdfad chromatin states. The width of columns corresponds to the proportion of chromatin states in H1hesc. The segmentation for the i^{th} column follows the proportion of windows in state ECi in H1hec that become Cj in Nhdfad. In other words, if we take the first pink rectangle of the first column, its width is proportional to the probability for a 100kb window to be in chromatin state EC1 in H1hesc and its height is proportional to the the probability for a 100kb window to be in C1 in Nhdfad given that it is in EC1 in H1hesc. The area of this rectangle (product of the previously mentioned probability) is proportional to the probability for a window to be in state EC1 in ESC and C1 in Nhdfad. (B) Same as (A) for the chromatin state changes from the cell line Nhdfad towards K562.

contain serum, Nhdfad medium as well as the three hematopoetic cell media did, which might explain the rather large C4 coverages obtained for K562 (26%) and Monocd14ro1746 (30%). Note that despite the fact that they were grown in similar conditions, Gm12878 cells have a significantly lower C4 coverage (11%). However, this seems to be compensated by a very high C3 coverage (36%) so that, as in the other differentiated cell lines, the total (C3+C4) coverage is $\sim$ 45% (Table V.1). This genome over-coverage by the unmarked silent C3 state in Gm12878 may be a particularity of lymphoid derived cells that remains to be understood.

To study changes in chromatin states between different cell lines, we selected two representative differentiated cell lines, namely Nhdfad and K562,

and the pluripotent H1hesc cell line. As an illustration of the transition between ESCs and differentiated cells, the changes obtained from H1hesc chromatin states to Nhdfad chromatin states (Fig. V.6A) are very instructive. The transcriptionally active state is highly conserved: 80% of EC1 100 kb-loci in H1hesc are C1 loci in Nhdfad as compared to 13% that experience a repression by polycomb to C2 and only 4% and 3% that transit towards the heterochromatin states C3 and C4 respectively. The bivalent state EC2 directs towards either the active euchromatin state C1 (29%) or the polycomb repressed state C2 (51%) which is coherent with initial bivalency adding flexibility in transcription regulation during development [6,8,15,150,162,164,166,228,243,269]. The unmarked state EC3 mainly leads to the heterochromatin states C3 (30%) and C4 (51%) and almost never to the active state C1 (5%). EC4 does not change much to the active state C1 (7%) but distributes almost equally into C2 (34%), C3 (21%) and C4 (39%). Even though they are quite different in terms of epigenetic marks (Figs V.4 and V.5), these three states are silent [154, 193]. Therefore EC4 state in pluripotent cells is prepared to silencing during differentiation. Now when looking at chromatin state changes from differentiated cell lines Nhdfad to K562 (Fig. V.6B), we observed that a majority of 100-kb loci in C1 (72%), C3 (58%), C4 (66%) and to a lesser extend C2 (33%) are conserved. Indeed except the highest percentage of C2 loci leading to C3+C4 (26%), the drastic difference is that the constitutive heterochromatin state C4 rarely transits to the active euchromatin state C1 (4%) and to the polycomb repressed state C2 (15%), which confirms that the pluripotent state EC4, if prepared to silencing, is not as C4, a compactly repressed heterochromatin state. Note that overall, chromatin states are highly dynamic since only 48% (resp. 57%) of 100 kb-loci are conserved from H1hesc (resp. Nhdfad) to Nhdfad (resp. K562). Merging the genic chromatin states EC1+EC2 (resp. C1+C2) significatively increases the conservation rate to 83% (resp. 69%). The merging of EC3+EC4 (resp C3+C4) also displays high conservation rate 74% (resp. 89%).

V.2.4 Replication timing of chromatin states

Consistent with our preliminary analysis of the K562 cell line in Chapter III [154], we confirmed that there exists a strong correlation between the four prevalent chromatin states and the MRT, and this for both the pluripotent (H1hesc) and the differentiated (K562, Gm12878, Nhdfad) cell lines (Fig. V.7).

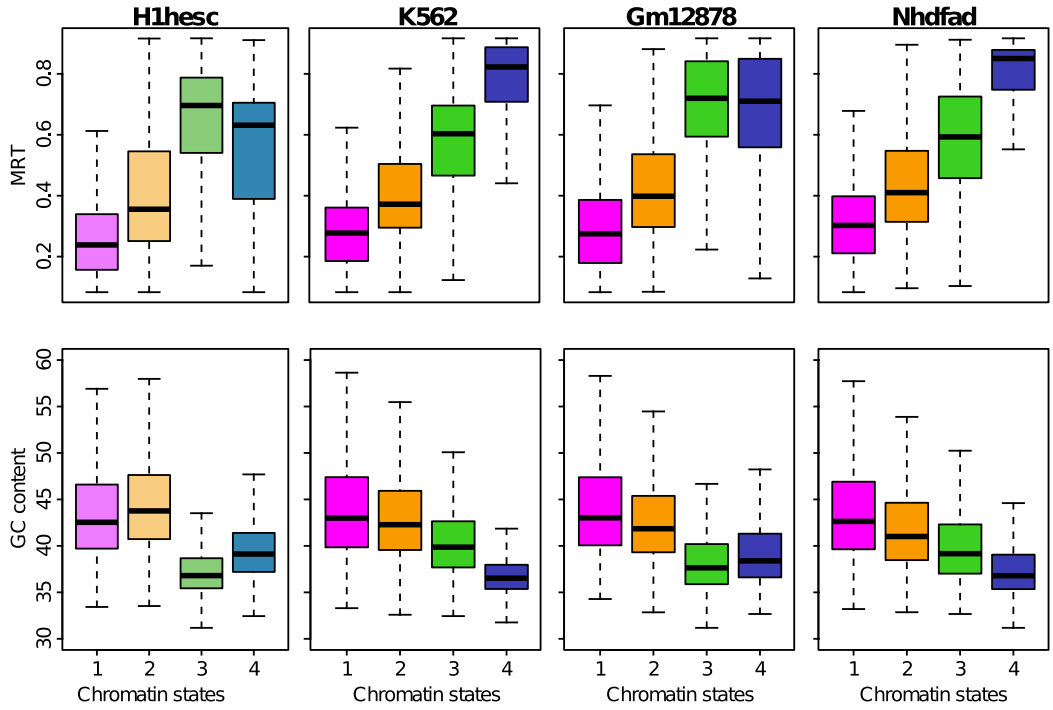

Figure V.7: MRT and GC distributions in the four chromatin states for h1Hesc and three differentiated cell lines (K562, Gm12878, Nhdfad). (First row) Boxplot of MRT computed in 100 kb non-overlapping windows per chromatin state. (Second row) Boxplots of GC content computed in 100 kb non-overlapping windows per chromatin state. Same color coding as in Fig. V.4.

The transcripionally active euchromatin states C1 and EC1 replicate early in S-phase in agreement with previous studies of open chromatin marks in human and mouse [12, 14, 34, 36, 87, 89]. The pluripotent bivalent EC2 state and the differentiated polycomb repressed C2 heterochromatin state both replicate slightly later in mid-S phase which contrasts with previous report of the existence of high correlation between late replication and the repressive chromatin mark H3K27me3 [36, 181]. The silenced unmarked EC3 and C3 states as well as the pluripotent chromatin state EC4 prepared to heterochromatinization and the HP1-associated heterochromatin state C4 all replicate much latter up to the end of S-phase. Interestingly, whereas (EC1, C1) and (EC2, C2) have clear different MRT, they have almost the same high mean GC content as expected for gene-rich states [1]. In contrast, a clear correlation between MRT and mean GC content was observed for the late replicating chromatin states. When C3 replicates before C4 (K562, Nhdfad), C3 has a higher GC con-

	from H1hesc to Nhdfad		from Nhdfad to K562	
	E to L	L to E	E to L	L to E
1 to 1	0.35	0.49	0.81	0.37
1 to 2	1.26	0.46	2.11	0.22
1 to 3	3.52	0.36	4.12	0.04
1 to 4	4.33	0.10	4.90	0.18
2 to 1	0.21	1.61	0.34	1.71
2 to 2	0.80	1.18	1.36	0.98
2 to 3	2.16	1.26	3.17	0.26
2 to 4	1.86	0.10	3.26	0.19
3 to 1	0.17	3.63	0.30	4.14
3 to 2	0.45	3.14	0.84	3.26
3 to 3	0.89	1.28	1.68	0.90
3 to 4	0.65	0.23	1.35	0.25
4 to 1	0.41	3.49	0.09	4.58
4 to 2	0.65	3.04	0.13	3.95
4 to 3	1.45	1.49	0.35	1.13
4 to 4	1.34	0.43	0.22	0.14

Table V.2: EtoL and LtoE chromatin state transitions from H1hesc to Nhdfad and from Nhdfad to K562. Chromatin state transition ratio is defined as the proportion of EtoL (resp. LtoE) transitions from state i to j over the proportion of this chromatin transition genome wide. Early (resp. late) means MRT < 0.5 (resp. > 0.5).

tent than C4 and vice-versa when C3 replicates after C4 (H1hesc, Gm12878) (Fig. V.7). There is however a major difference between MRT of pluripotent and differentiated cell lines. EC4 exhibits a much wider MRT distribution than C4 with a non-negligible proportion of early replicating (MRT < 0.5) 100-kb loci, namely 35.7% (H1hesc) as compared to 5.5% (K562), 19.2% (Gm12878) and 4.2% (Nhdfad). This can be seen as an additional indication that EC4 is sufficiently accessible and open to enable origin firing and early replication. This is confirmed by the almost uniform distribution of DNaseI hypersensitive sites (DHS) in H1hesc EC1 (median at 5 DHS per kb), EC2 (5 DHS/kb) and EC4 (1 DHS/kb) which contrasts with the abundance of DHS in differentiated C1 (40 DHS/kb) and C2 (20 DHS/kb) states and their virtual absence in the heterochromatin states C3 and C4 (Fig. V.4). The highly dynamic and accessible character of pluripotent chromatin states likely facilitates the access of the replication machinery to DNA and thus prevents having to repli-

cate long (EC3+EC4) threads at the end of S-phase. Replication of several Mb long silent fragments at the end of S-phase may be in contradiction with the necessity of maintaining a high proliferation rate and a short cell cycle in H1hesc [270].

To quantify the coupling between chromatin state transitions and MRT changes, we investigated the types of chromatin state transitions in EtoL and LtoE 100 kb windows that were over-represented as compared to their genome wide proportions (Table V.2). For the EtoL transitions from H1hesc to Nhdfad, besides the polycomb repressed EC1 to C2 transition, the heterochromatization transitions from (EC1, EC2) to (C3, C4) are as expected significantly over-represented and account for 33% of all EtoL transitions (10.8% of transitions genome-wide). Interestingly, there are more surprising EtoL chromatin state transitions from EC4 to C3 and C4. They correspond to the early replicating EC4 100 kb loci that "consolidate" into silent and compact heterochromatin loci replicating late in Nhdfad. These transitions account for 30% of all EtoL transitions (17% genome wide) as compared to 60% from (EC1, EC2) to (C3, C4) EtoL transitions. For the opposite LtoE transitions from H1hesc to Nhdfad, besides the expected transitions from the bivalent EC2 state to the active C1 state and to the polycomb repressed C2 state, the activation transitions from (EC3, EC4) to C1 are highly over-represented (8%) (2% genome-wide) as well as transitions from (EC3, EC4) to the polycomb repressed C2 state (40%) (12% genome wide). Between differentiated cells, same trends were obtained for EtoL and LtoE chromatin state transitions from Nhdfad to K562 with the remarkable difference that the over representations of EtoL transitions from C4 to C3 and to C4 are no more present as the consequence of the absence of early initiation in compact HP1-associated heterochromatoin domains (Table V.2, Fig. V.7)

V.2.5 Gene content of chromatin states

To address the issue of gene content of pluripotent and differentiated prevalent chromatin states, we focused on H1hesc and K562 cell lines. We took advantage of our previous detailed integrative analysis of epigenetic marks, MRT and gene expression data in K562 in Chapters III and IV [154,193] that showed that the euchromatin state C1 is highly genic and contains almost all expressed genes and a non negligible proportion of inactive genes that almost equals the total number of genes found in C2 as mostly repressed by polycomb complexes. As

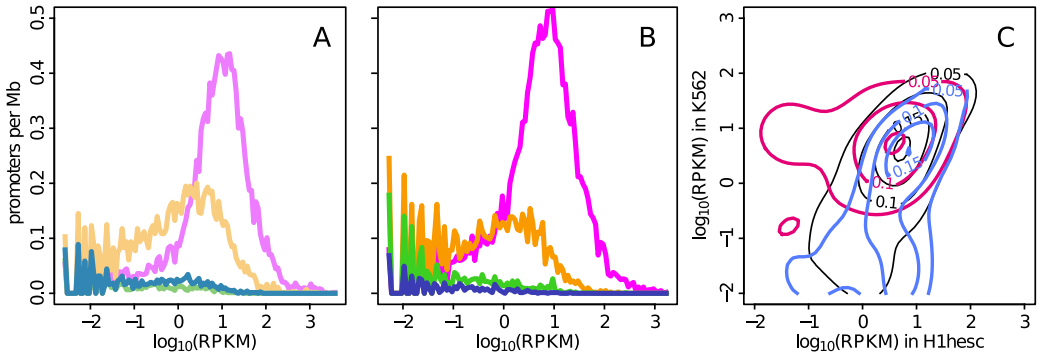

Figure V.8: Gene expression in the H1hesc and in the K562 chromatin state. (A) Density of promoters in the 4 chromatin states of the H1hesc cell line as a function of gene expression (genes were grouped into bins of width 0.05 in $\log_{10}$(RPKM) unit). Same color coding as in Fig. V.3D. (B) Density of promoters in the 4 chromatin states of the K562 cell line as a function of gene expression. Same color coding as in Fig. V.3D'. (C) 2D representation of the joint density of gene expression in H1hesc (X-axis) and K562 (Y-axis) when focusing on EtoL (blue) and LtoE (magenta) MRT transitions. For comparison is shown as a control (black), the joint density obtained for comparable size sets of randomly chosen genes.

compared to these high-GC (Fig. V.7), gene rich C1 and C2 states, the low-GC C3 and C4 states were found to be gene deserts with scarce long genes. In the pluripotent H1hesc cell line, the gene rich chromatin states are still EC1 and EC2. But there are some noticeable differences with respect to K562. There are less promoters per Mb in EC1 (13.1 promoters/Mb) than in C1 (15.9 promoters/Mb), and in compensation more in EC2 (9.3 promoters/Mb) than in C2 (7.8 promoters/Mb) (Table V.3). Moreover the relative distributions of RPKM values (Eq. (V.4))(Fig. V.8A,B) revealed that relative to C1, EC1 contains more expressed genes with RPKM > 1 as well as EC2 relative to C2. Indeed, both mean and median RPKM values are higher in EC1 and EC2 than in C1 and C2 respectively (Table V.3). This is consistent with the extensive presence of bivalent genes in EC2 that was previously shown to be more accessible and less compact than the polycomb repressed C2 state in differentiated cell lines [238,269]. This is also in agreement with previous report on the higher global transcription activity in ESCs with only sporadic tissue-specific gene expression as compared to differentiated cells [271]. Note that, in that respect, EC4 is slightly permissive to expression whereas C4 is the most repressive heterochromatin state (only 25% of genes with a non-null RPKM) with by far the lowest gene density and largest mean length (Table V.3).

Chromatin states	EC1	EC2	EC3	EC4	C1	C2	C3	C4
promoter count	7653	5047	1138	1505	9588	3089	2334	827
promoter density per Mb	13.127	9.278	1.497	2.16	15.861	7.812	2.953	1.133
mean gene expression (RPKM)	30.537	7.76	0.798	1.299	18.529	2.307	0.696	0.244
median gene expression (RPKM)	8.527	0.988	0.008	0.025	5.58	0.29	0.005	0
mean gene length (kb)	50.999	72.705	56.335	67.538	43.557	58.535	86.301	173.962

Table V.3: Gene content in the four prevalent chromatin states of H1hesc and K562 cell lines. For each chromatin state the following information is given: (i) the total number of promoters per chromatin state, (ii) the density of promoters per Mb, (iii) the mean level of expression per chromatin state in RPKM (Eq. (V.4)), (iv) the median level of expression per chromatin state in RPKM, (v) the mean gene length per chromatin state in kb.

The coupling between MRT and gene expression has been extensively studied in *Drosophila* [84, 85, 88] and mammals [11, 12, 34, 86, 217]. We found that in both H1hesc and K562, a vast majority of expressed genes are in the early replicating EC1 and C1 chromatin states which confirms the link between MRT and expressed gene density previously reported in mouse [34, 35, 86] and human [11, 12, 14, 154]. Even more, in Chapter IV [193] we showed that the activation of one gene in K562 was almost always sufficient for its 100 kb environment to be in a early C1 chromatin state. But the presence of an important number of inactive genes in early C1 regions and to a less extend in early EC1 regions (Table V.3, Fig. V.8A,B), suggests that there is no causal link between an early replicating region and a high expression level yet many recently identified early replication origins are strongly associated with CGI and active CpG-rich gene promoters [56, 61, 70, 71, 73, 74, 79, 80, 272]. If almost all genes in the late replicating heterochromatic C3 and C4 states are silent with few exceptions, there is a slighlty larger number of expressed genes in the pluripotent EC4 state (25% of the few genes in EC4 100kb windows have a non null RPKM). Recent studies in mammals have further shown that the dynamic of MRT through differentiation is only loosely coupled with gene expression dynamic [12, 34, 35, 88]. Whereas in dynamic timing regions most genes do not change expression [12], some genes can undergo silencing while being replicated early [12, 34, 35]. When examining the joint distribution of gene expression in H1hesc and K562 (Fig. V.8C), for 100kb loci that experience a EtoL transition and reversely for those that change from LtoE, we confirmed that most ($\sim 55\%$) genes lying in dynamic MRT regions do not change expression. Only a small fraction of genes that fall in the tails of the joint distribution change

Chromatin state	EC1	EC2	EC3	EC4	EC1+EC2	EC3+EC4
H1hesc	296.14	241.89	398.56	300.81	569.34	740.2
Chromatin state	C1	C2	C3	C4	C1+C2	C3+C4
K562	327.43	191.14	437.69	881.56	567.19	869.6
Monocd14ro1746	357.48	215.22	210.33	582.15	642.93	628.52
Gm12878	336.87	198.34	576.08	276.3	551.83	702.88
Hmec	322.34	231.96	357.57	610.08	608.46	563.51
Nhdfad	312.11	261.89	281.14	894.38	675.45	637.58

Table V.4: Mean length of chromatin state blocks per chromatin state in kb (Sect. V.5.13) in ES H1hesc and differentiated cells (see Table V.1).

expression coordinately with MRT change (Fig. V.8C), EtoL corresponding to HP1-associated heterochromatin silencing (5.3%) and LtoE to open chromatin activation (13%). We considered that genes change expression if their expression was at least three fold different between the considered cell lines. As discussed in [12,34,35], genes that have the ability to overcome heterochromatin repression during a EtoL MRT change are mainly genes that have a strong CpG rich promoter (*e.g* housekeeping genes). On the other hand, gene activation seems to concern CpG rich as well as CpG poor promoters suggesting that a switch LtoE generates a permissive environment to transcription.

V.2.6 Spatial organization of chromatin states along human chromosomes

Once mapped on the genome (Fig. V.9), the organization of the four prevalent chromatin states looks quite different in the pluripotent H1hesc cell line as compared to the one in the five differentiated cell lines (Table V.4). In H1hesc, the four chromatin states EC1, EC2, EC3 and EC4 do not differ so much in the genome coverage (Table V.1) as well as in the number and length distributions of domains or blocks of adjacent 100-kb-loci in the same chromatin state (Table V.4, Fig. V.10A). In Nhdfad, in agreement with previous analysis in K562 in Chapter III [154], the HP1-associated heterochromatin state C4 has a length distribution that displays a fat tail not observed in the C1, C2 and C3 length distributions (Fig. V.11A) as well as in the corresponding H1hesc length distributions (Fig. V.10A). This fat tail explains that the mean C4 block length ($\bar{\mathrm{L}} = 894$ kb) is significantly larger than the mean block

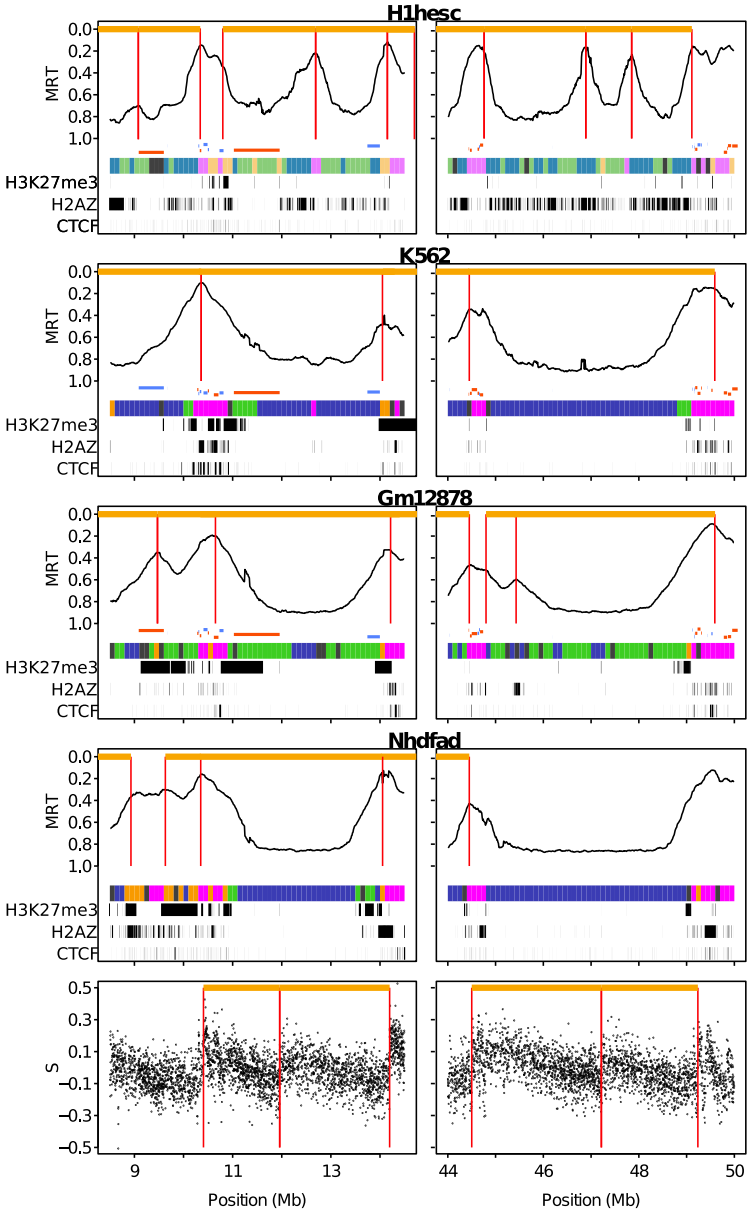

Figure V.9: Genome-wide spatial distribution of chromatin states in ESCs and differentiated cells. MRT profile along two Mb long fragments of human chromosome 5. U-domains are marked by an horizontal orange line and their borders by vertical red lines. Below the MRT profile, gene positions are indicated by a horizontal segment (blue: not expressed, orange: expressed) as well as the chromatin state of each 100 kb window is represented using the same color coding as in Fig. V.3D,E. At the bottom of the plot, intervals significantly enriched in H3K27me3, H2AZ and CTCF are represented in black. At the bottom of the figure, the last panel represents the skew $S = S_{GC} + S_{TA}$ with germline replication skew N-domains are marked by an horizontal orange line and their borders by a vertical red line.

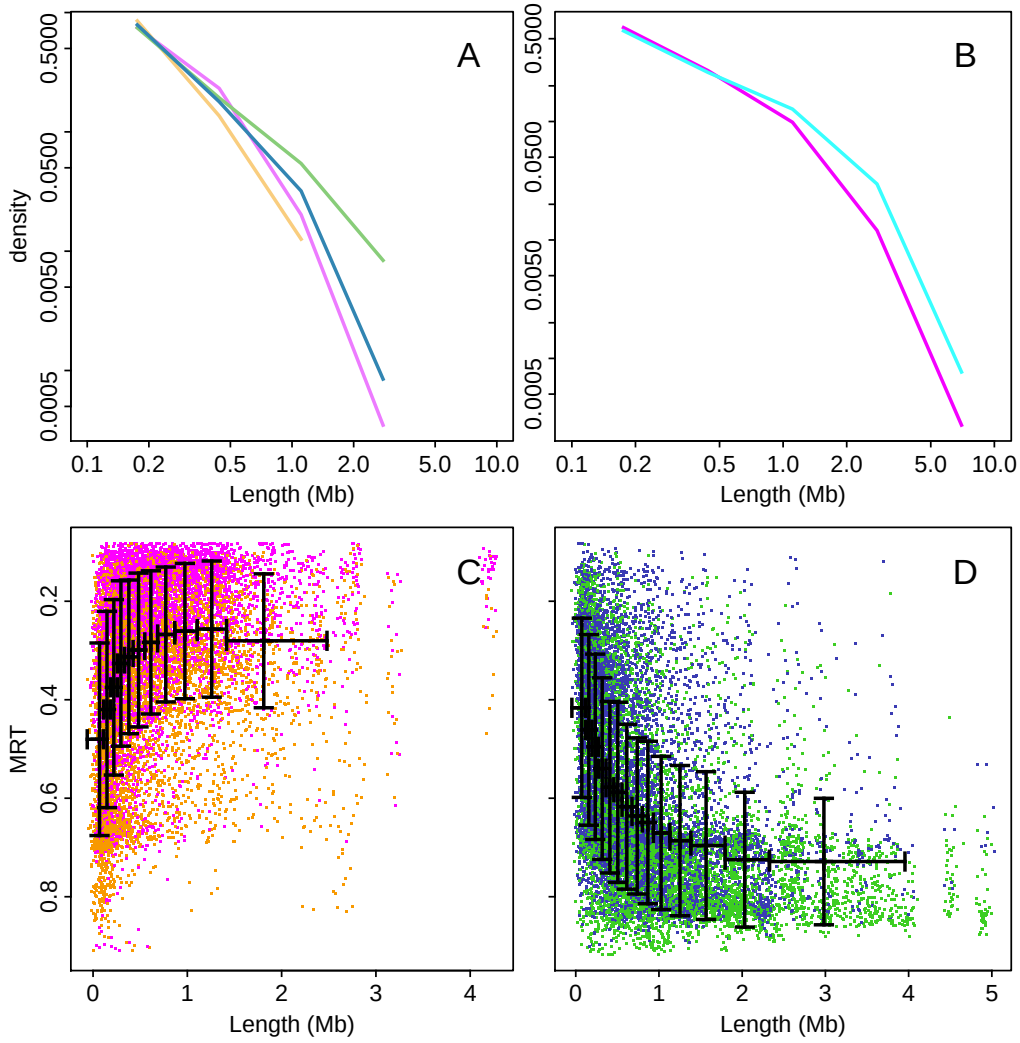

Figure V.10: Spatial organization of chromatin states in H1hesc. (A) Histogram of chromatin state block length in a logarithmic representation (Sect. V.5.13). (B) same as (A) for chromatin blocks formed by states EC1 and EC2 (EC1+EC2, light red) or by states EC3 and EC4 (EC3+EC4, light blue). (C) MRT in chromatin state blocks EC1+EC2 with respect to their length. Each 100 kb window in a chromatin state block is represented by the color of its state defined in Fig. V.3D,E. The mean profile was obtained by (i) ordering data points according to their block length, (ii) grouping them in classes of equal number of data points and (iii) computing the average length and MRT over each class. Vertical bars represent the standard deviation. Horizontal bars represent the range of length over each class. (D) Same as (C) for chromatin state blocks EC3+EC4.

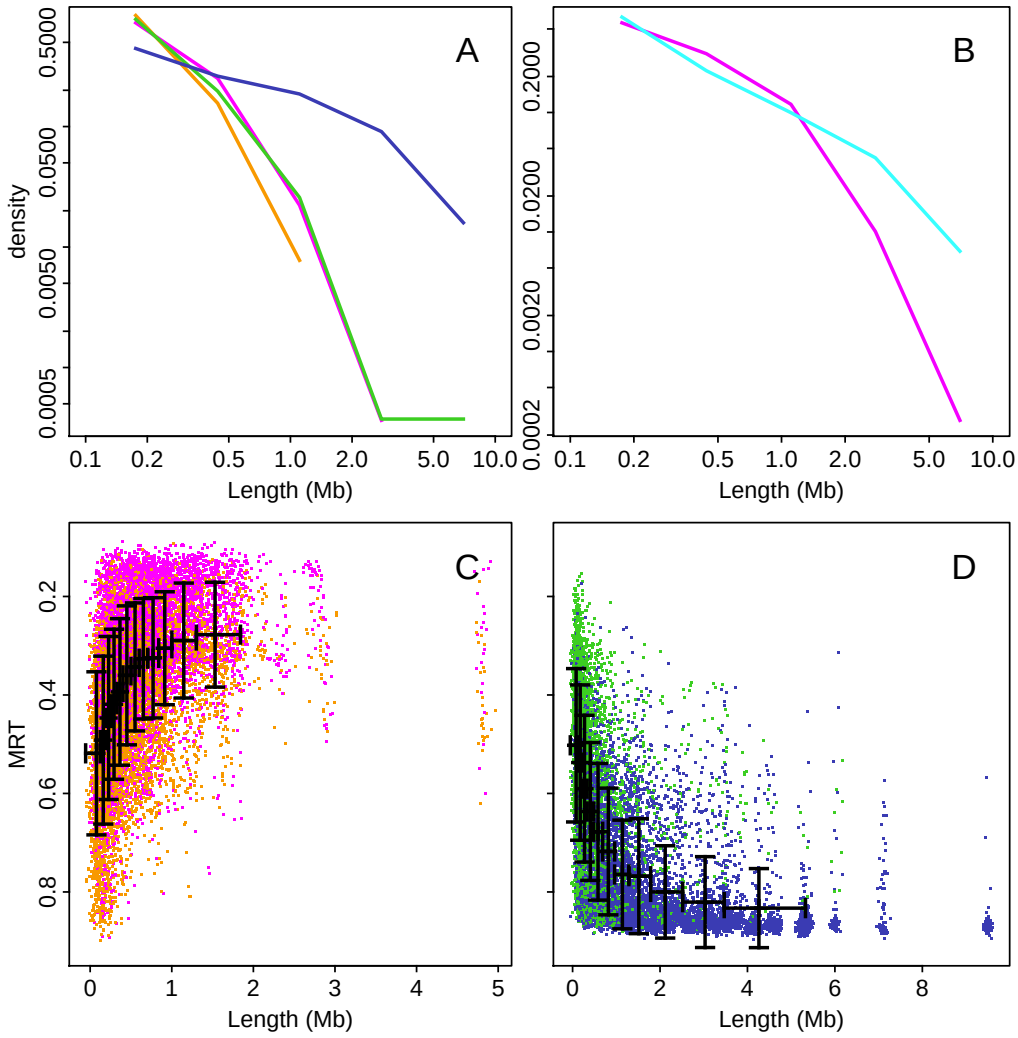

Figure V.11: Spatial organization of chromatin states in Nhdfad. (A) Histogram of chromatin state block length in a logarithmic representation (Sect. V.5.13). (B) same as (A) for chromatin states C1 and C2 (C1+C2, light red) or by states C3 and C4 (C3+C4, light blue). (C) MRT in chromatin state blocks C1+C2 with respect to their length. Each 100 kb window in a chromatin state block is represented by the color of its state defined in Fig. V.3D',E'. The mean profile was obtained by (i) ordering data points according to their block length, (ii) grouping them in classes of equal number of data points and (iii) computing the average length and MRT over each class. Vertical bars represent the standard deviation. Horizontal bars represent the range of length over each class. (D) Same as (C) for chromatin state blocks C3+C4.

length of C1 ($\bar{L}$ = 312 kb), C2 ($\bar{L}$ = 262 kb) and C3 ($\bar{L}$ = 281 kb). This peculiar length property of C4 blocks is shared by all differentiated cell lines except Gm12878 where C3 blocks are larger ($\bar{L}$ = 576 kb) as compared to C4 blocks ($\bar{L}$ = 276 kb). Interestingly, as originally observed in K562 [154], for all differentiated cell lines as well as for the ESC line H1hesc, the association of C1+C2 (resp. EC1+EC2) on one side and C3+C4 (resp. EC3+EC4) on the other side, results in large scale blocks of surprisingly similar length distributions (Table V.4, Fig. V.10B and V.11B). But the length distributions obtained for differentiated cells have a fat tail up to blocks larger than 10 Mb (Fig. V.11B and also Fig. III.19D) whereas EC1+EC2 and EC3+EC4 blocks in H1hesc do not exceed 5 Mb (Fig. V.10B). These very long C1+C2 blocks actually replicate very early (Fig. V.11C) suggesting that C2 loci are replicated passively from fork coming from neighboring active loci earlier than C2 loci isolated in a C3, C4 environment. On the contrary, very long C3+C4 blocks definitively replicate very late (Fig. V.11D) as expected for gene desert low-GC heterochromatin regions. These results are quite consistent with the statistical model proposed in [12] where MRT is predicted from the distance to the nearest active promoter. In H1hesc, the long EC1+EC2 (resp. EC3+EC4) blocks also correspond to early (resp. late) replicating regions (Figs V.10C and D) which explains that because of a shorter cell cycle in ESC, their maximal length ($\sim$ 5Mb) is significantly shorter than in differentiated cells ($\sim$ 12 Mb). Furthermore when looking at the most dynamical parts of the human genome in terms of terms of MRT, we found that those that switch MRT from EtoL or LtoE are small fragments of a few 100 kb long (Fig. V.12). From H1hesc to Nhdfad, very small EC1+EC2 fragments switch to later MRT and mainly correspond to the polycomb repressed EC1 to C2 transition and to the heterochromatization transitions (EC1, EC2) to (C3, C4) (Table V.2, Fig. V.12A). Also some small EC1+EC2 fragments switch to earlier MRT and in particular the expected transition from the bivalent EC2 state to the active C1 state. Similarly, only small EC3+EC4 fragments switch to earlier MRT and actually correspond to the activation transitions from (EC3, EC4) to C1 and to the transitions (EC3, EC4) to C2 (Table V.2, Fig. V.12B). Some small EC3+EC4 fragments switch to much later MRT as well as some individual 100 kb loci belonging to larger EC3+EC4 fragments. They correspond to the few early EC4 100 kb loci in H1hesc that consolidate into late replicating silent (C3) and compact HP1-associated (C4) heterochromatin loci (Table V.2). From Nhdfad to K562 similar trends were observed with only small C1+C2 and C3+C4 frag-

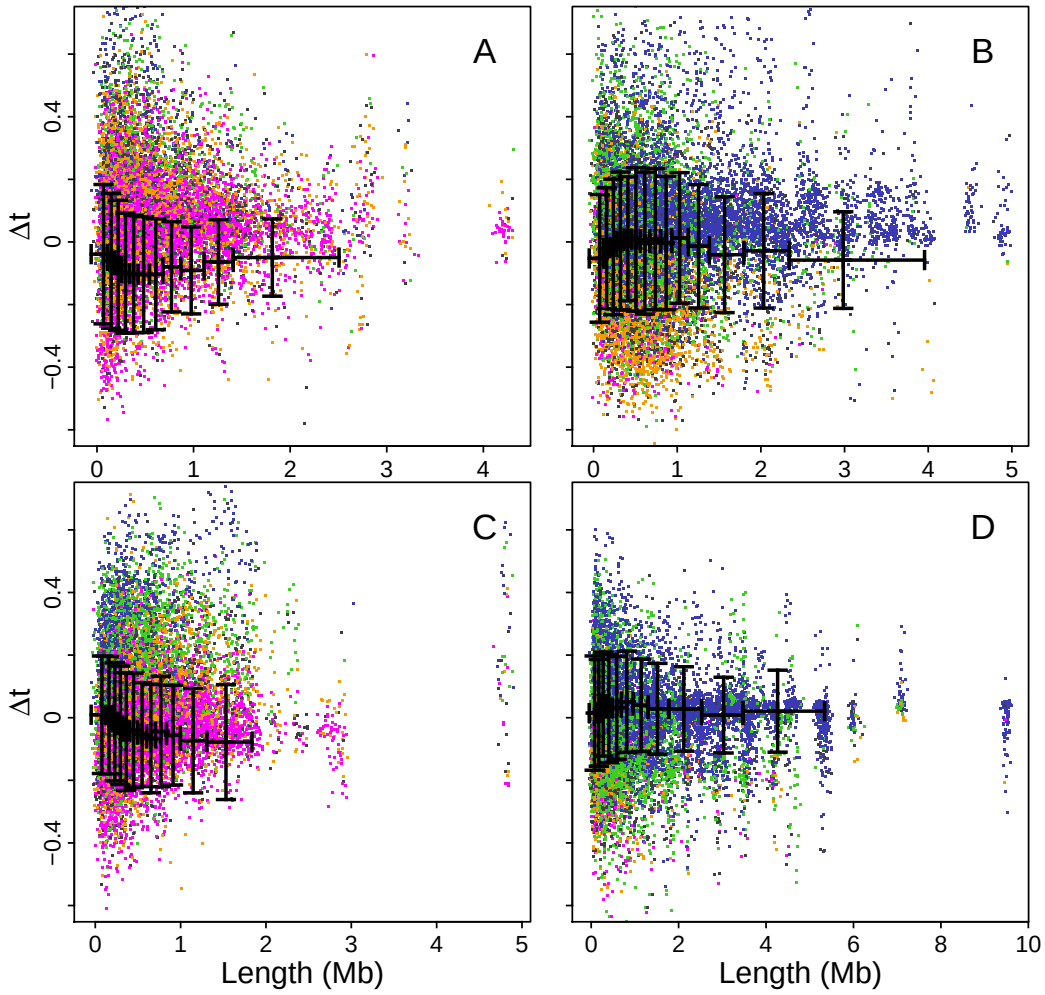

Figure V.12: (A) MRT changes between H1hesc and Nhdfad ($\Delta_t = \mathrm{MRT}_{Nhdfad} - \mathrm{MRT}_{ESC}$) in chromatin blocks EC1+EC2 with respect to their length. Each 100 kb window in a chromatin state block is represented by a dot of its corresponding chromatin state color in Nhdaf. (B) same as (A) for chromatin blocks C3+C4. (C) same as (A) for the cell lines Nhdfad and K562. (D) same as (B) for the cell lines Nhdfad and K562.

ments changing MRT except that, as previously noticed, there was no more shift towards later MRT in long C3+C4 fragments due to the absence of early initiation in these heterochromatin domains (Table V.2, Fig. V.12C and D). Altogether these results confirm that the genomic context is important to MRT and chromatin state dynamic.

V.2.7 Distributions of chromatin states inside and outside replication U/N-domains

In all the cell lines examined in this work, about half of the human genome was shown to be paved by replication timing U-domains (Fig. V.9) [94]. However their number (N) and mean length ($\bar{L}$) drastically differ in H1hesc (N= 1534, $\bar{L}$= 1.09 Mb) and in the differentiated cell lines K562 (N=876, $\bar{L}$=1.42 Mb), Gm12878 (N=882, $\bar{L}$=1.52 Mb) and Nhdfad (N=1150, $\bar{L}$=1.19 Mb). MRT U-domains are more numerous and shorter in the ESC line than in the differentiated cell lines as the probable consequence of the shorter S-phase duration in pluripotent cells. Interestingly the number of U-domains shared by each cell type pairs including the germline replication skew N-domains, was shown to be significantly larger than the number expected by chance [94] which put forward these replication U/N-domains as robust features of the spatio-temporal replication program in human.

Chromatin state organization inside replication U-domains

When concentrating our study on the replication U-domains identified in H1hesc (Fig. V.13A-D) and Nhdfad as a representative of differentiated cell lines (Fig. V.13A'-D'), we revealed some remarkable organization of the four prevalent chromatin states with some notable differences that distinguish the global dynamical and accessible character of pluripotent chromatin from the expanding HP1-associated heterochromatin in differentiated cells. Consistent with the organization found in K562 (Fig. III.20) [154], the highly expressed gene-rich open euchromatin state C1 was found to be confined in a closed ($\lesssim$ 150 kb) neighborhood of the master replication origins that border each individual U-domains (Fig. V.13A') and this independently of the domain size. Significantly enriched in DHS and CTCF (Fig. V.4), C1 can thus be seen as specifying the early initiation zones that border U-domains and that were further shown to delimit topological domains on genome-wide (4C, Hi-C) chromatin state

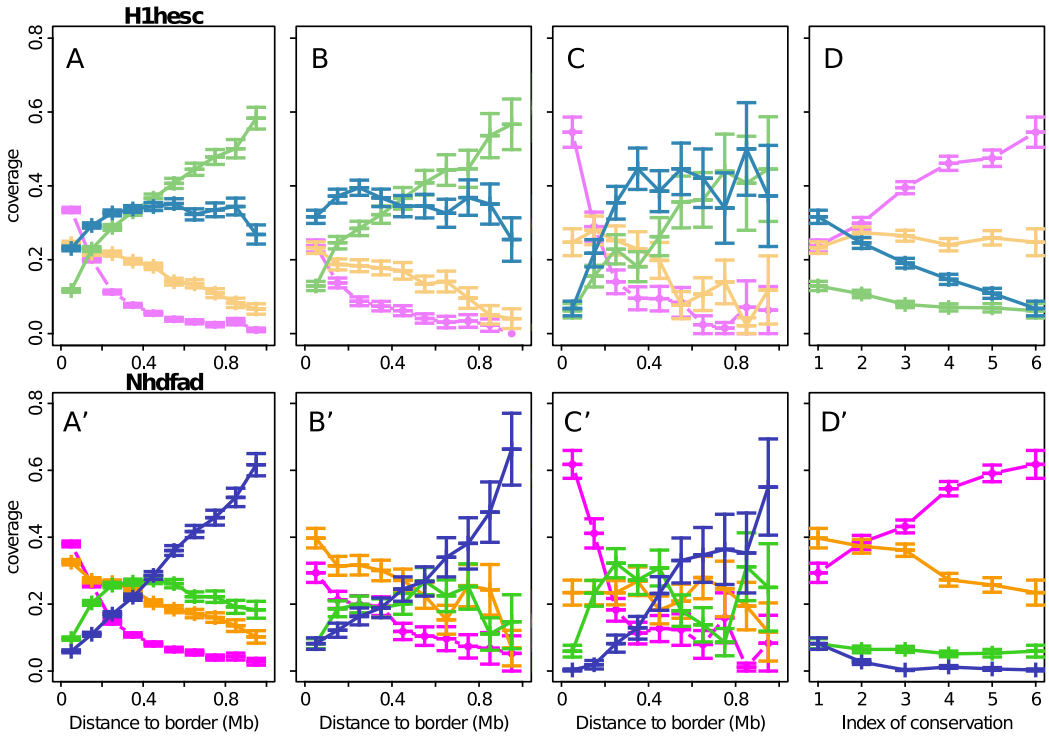

Figure V.13: Distribution of chromatin states inside replication timing U-domains of H1hesc and Nhdfad. (A) Mean coverage of chromatin states with respect to the distance to the closest U-domain border in the H1hesc cell line. (B) Same as (A) for U-domains specific to the H1hesc cell line. (C) same as (A) for H1hesc U-domains conserved in all cell lines. (D) Mean coverage of ESC chromatin state in the 100kb window containing a U-domain border with respect to the conservation index n of the U-domain border (Sect. V.5.15). (A'-D') same as (A-D) for the Nhdfad cell line. Same color coding as in Fig. V.3D and V.3D' respectively.

conformation data [30, 94, 113]. The polycomb repressed state C2 was mainly found occupying the mid-S phase 200-300 kb region away from U-domain borders (Fig. III.20). Remarkably, U-domain borders are significantly depleted in unmarked (C3) and constitutive (C4) heterochromatin states (Fig. V.13A'). C3 is present in the center of small U-domain and homogeneously occupies large U-domain centers. C4 is abundantly found in the center of large U-domains ($\gtrsim$ 1Mb). These results for Nhdfad and K562 (Fig. III.20) [154] confirm that the replication "wave" starting from the early initiation zones at U-domain borders and propagating inside these domains via the progressive activation of secondary origins [93, 109], actually progress in a gradient of chromatin structures from openess (C1) to compactness (C3, C4), via the polycomb repressed

Chromatin state	EC1	EC2	EC3	EC4	EC1+EC2	EC3+EC4
H1hesc	203.28	163.41	235.97	196.98	308.43	373.07
Chromatin state	C1	C2	C3	C4	C1+C2	C3+C4
K562	209.38	130.94	270.62	718.53	292.41	516.22
Gm12878	223.54	142.19	339.2	211.25	299.62	382.95
Nhdfad	213.56	168.54	197.61	654.68	328.76	403.81

Table V.5: Distribution of chromatin states outside replication timing U-domains. Same as Table V.4 after removing the replication U-domains from the analysis.

state C2.

In the smaller H1hesc U-domains, the concentration of EC1 around the bordering master replication initiation zones and the distribution of EC2 nearby in mid-S phase proximal regions (Fig. V.13A) ressembles to the organization of high-GC, gene-rich chromatin states (C1, C2) in differentiated cells (Fig. V.13A'). However the distributions of EC3 and EC4 (Fig. V.13A) are drastically different from those of C3 and C4 in Nhdfad (Fig. V.13A') and K562 (Fig. III.20) [154]. EC3 is still depleted at U-domain borders and mainly covers the center of the largest U-domains. Importantly, unlike C4, EC4 is now abundantly found at U-domain borders as well as inside these domains. As addressed in the Sect. V.3, this homogeneous distribution of the gene-poor silent EC4 state inside replication U-domains actually reflects the almost uniform covering (inside U-domains as well as outside) of the human genome by the histone variant H2AZ in pluripotent cells (Fig. V.14A) [238].

Chromatin state organization outside replication U-domains

The complete analysis of the other half of the genome that is complementary to U-domains was more in agreement with the dichotomic view proposed in early studies of the mouse [34, 35, 86] and human [12, 36, 90] genomes, where early and late replicating regions occur in separated compartments of open and close chromatin, respectively. About 25% of the human genome are covered by megabase sized gene-rich, high-GC EC1+EC2 (resp. C1+C2) chromatin blocks in H1hesc (resp. differentiated) cells (Table V.5), that on average replicate early (Figs V.10C and V.11C) by multiple almost synchronous origins with equal proportion of forks coming from both directions. Since the replication fork polarity is reflected in the MRT derivative [94, 107], these regions

correspond to early MRT plateaus that are remarkably well conserved between pluripotent and differentiated cell lines. The last 25% of the human genome correspond to megabase sized gene-poor, low-GC EC3+EC4 (resp. C3+C4) chromatin blocks in H1hesc (resp. differentiated) cells (Table V.5), that on average replicate late by again multiple almost coordinated origins. In regard to the drastic difference in chromatin properties of the silent states EC4 (dynamically accessible) and C4 (compact heterochromatin), these late MRT plateaus must simply be seen as the late replicating counter part of the early (C1+C2) plateaus.

V.3 Discussion

V.3.1 Specific genome-wide histone signature of pluripotent plastic chromatin

Our integrative analysis of epigenetic marks confirmed the existence of fundamental differences between the pluripotent and differentiated chromatin states (Fig. V.4). These differences account for the drastic changes observed in epigenetic landscapes in ESC and lineage committed cells (Fig. V.9) [6, 166, 228, 236, 238, 250, 251]. In general, histone modifications show two distinct types of spatial distributions: small localized peaks and large spreading domains. The histone variant H2AZ associated with nucleosome exchange and remodeling [6, 55, 166, 202, 235, 238, 257], was typically found confined to promoters and distal elements in differentiated cells [166, 238] which explains its abundance in the gene-rich chromatin states C1 and C2 (Fig. V.4). Its binding level was further shown to correlate with gene expression in human [166] which is consistent with its highest enrichment in the transcriptionally active state C1. Remarkably, the global H2AZ distribution diverges markedly between pluripotent and differentiation cells. In H1hesc, 92% of the overall 100-kb loci contain the H2AZ mark as compared to smaller coverages in K562 (61%), Gm12878 (65%), Hmec (76%) and Nhdfad (79%) (Note the important covering found for Monocd14ro (94%)) (Table V.6). Thus in ESCs, H2AZ marks promoters and distal elements but it is also distributed thoughout intergenic regions which explains its presence in the gene-rich chromatin states EC1 and EC2 and in addition, its specific abundance in the gene-poor chromatin state EC4 (Fig. V.4). This broad H2AZ distribution suggests that chromatin exchange

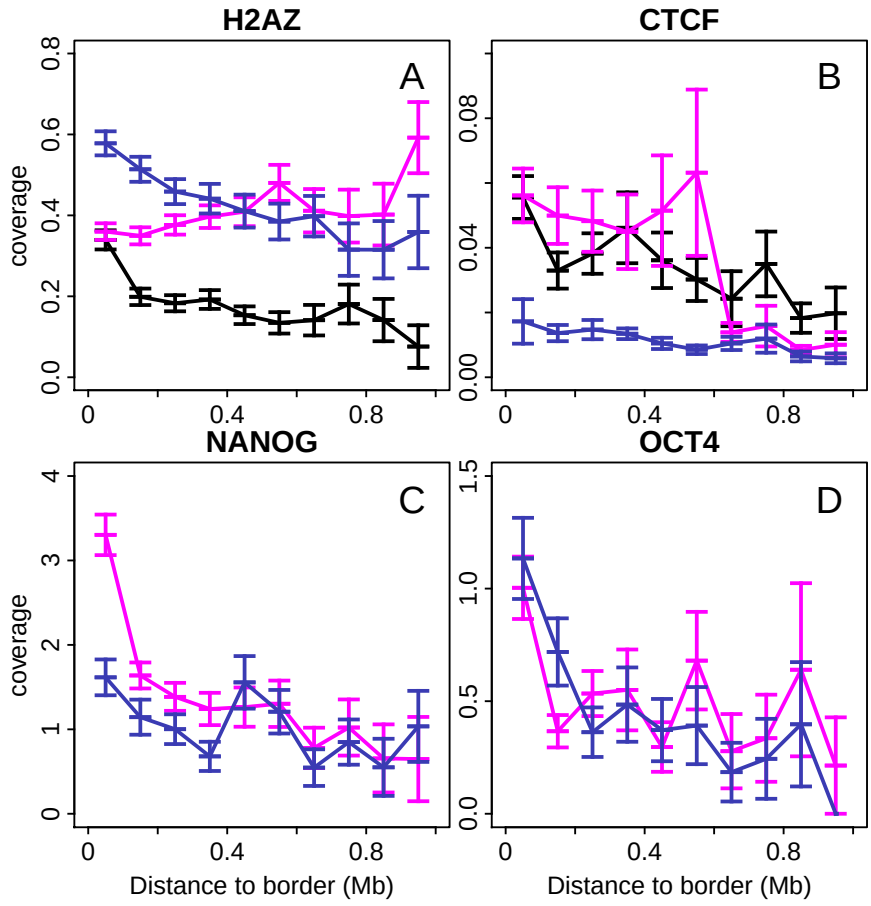

Figure V.14: Epigenetic marks enrichment in specific MRT U-domains of H1hesc and Nhdfad. (A) Mean coverage of H2AZ enriched intervals with respect to the distance to the closest U-domain border specific to the cell line. The different colors correspond to specific U-domains of Nhdfad (black), specific U-domains of H1hesc whose border is in EC1 or EC2 (red) and specific U-domains of H1hesc whose border is in EC4 (blue). (B), (C) and (D) are as (A) for respectively CTCF, NANOG anf OCT4.

and remodeling are prevalent throughout human chromosomes in ESCs [238]. This highly dynamic and potentially accessible properties of pluripotent chromatin are further strengthened by the presence of the ATP-dependent remodeler CHD1 not only in EC1 and EC2 but also in EC4 as an inhibitory factor to HP1-heterochromatin (Fig. V.4) [263]. In addition, we showed that the active marks H3K4me1 has a broad dispersion in H1hesc (85% coverage) relative to a more restrictive confinement at promoters and enhancers in differentiated cells K562 (55%), Monocd14ro1746 (61%), Gm12878 (62%), Hmec (77%) and Nhdfad (72%) (Table V.6) [238]. This is consistent with the abundance of H3K4me1 in the gene-rich EC1 and EC2 chromatin states and also with its presence in the gene-poor EC4 state contrasting its absence in the heterochromatin state C4 (Fig. V.4). There is another histone modification, namely H3K27me3, that distributes quite differently in pluripotent and differentiated cells. In ESCs, this surrogate of polycomb activity is mainly confined to "bivalent" promoters (37% coverage) that also carry H3K4me3. This contrasts with the much broader distribution of H3K27me3 in K562 (54%), Monocd14ro1746 (65%), Gm12878 (55%), Hmec (54%) and Nhdfad (60%). As indicated by the co-presence of H3K27me3 and H2AZ in the bivalent chromatin state EC2 (Fig. V.4) and the observed local surrounding of H3K27me3 marks by H2AZ variants in the H1hesc epigenetic landscape (Fig. V.9), the highly dynamic and potentially accessible pluripotent chromatin is likely refractory to polycomb facultative heterochromatin formation and spreading [228, 238]. The smaller mean size of EC4 blocks ($\bar{L}$ = 301 kb) in H1hesc as compared to C4 blocks in K562 ($\bar{L}$ = 882 kb), Hmec ($\bar{L}$ = 610 kb) and Nhdfad ($\bar{L}$ = 894 kb) (Table V.4), suggests that the gene-poor H2AZ marked accessible EC4 chromatin is incompatible with the stable interactions involved in the H3K9me3 enriched HP1 heterochromatin compaction and spreading (Fig. V.4). All the other histone marks known to be involved in transcription positive regulation, including H3K27ac and H3K36me3, have a similar distribution with a similar coverage of the gene-rich genome regions in H1hesc (EC1+EC2) and differentiated cells (C1+C2) (Table V.6).

V.3.2 Distinct epigenetic mechanisms of heterochromatin expansion during differentiation

There are mainly two epigenetic mechanisms of heterochromatin expansion during differentiation that correspond to the transitions towards the polycomb

	CTCF	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K9me3	H2AZ	H3K79me2	H4K20me1
H1hesc	88.15	63.65	36.48	39.76	84.67	75.14	48.53	81.61	91.58	36.55	50.41
K562	72.24	45.12	53.86	37.92	55.18	55.42	49.12	68.35	61.05	33.03	53.5
Monocd14ro	57.33	51.44	64.63	36.94	61.39	95.53	56.61	76.79	93.54	33.25	33.24
Gm12878	71.64	50.55	55.4	36.61	61.82	58.78	55.02	69.99	64.65	32.9	48.81
Hmec	72.27	59.57	54.11	39.41	76.85	74.32	49.18	57.25	76.11	36.22	56.94
Nhdfad	89.96	60.69	60.38	43.7	71.57	72.12	64.05	63.26	79.41	36.63	39.87

Table V.6: Genome coverage by epigenetic marks in ESCs and differentiated cells. Percentage of 100-kb windows that contain a given epigenetic mark.

repressed state C2 and towards the HP1-associated heterochromatin state C4 (Fig. V.6A). For the former mechanism, there are indeed two possible scenarios according to whether the pluripotent chromatin state that switches to C2 is EC2 or EC4. As previously described, EC2 is a bivalent chromatin state that is enriched in gene promoters that carry both the active mark H3K4me3 and the polycomb associated mark H3K27me3. This second mark has repressive effect on gene expression and contributes to maintain repression of bivalent genes including developmental genes [242–247, 255]. Some of these bivalent genes get activated during differentiation and switch from EC2 to the open euchromatin state C1 (Fig. V.6A). The other ones experience some repression to the facultative chromatin state C2 via the expansion of H3K27me3 to often cover the entire gene and frequently neighboring gene loci [228]. But there is another category of pluripotent genes that face this facultative heterochromatization which are the genes that are in the H2AZ rich accessible chromatin state EC4. These EC4 genes are actually lying nearby EC2 genes and get involved in the repressive expansion of H3K27me3 which dictates their switch to the polycomb repressed state C2. Note that this H3K27me3 spreading over several kb or tens of kb is locally at the expense of H2AZ which confirms that, in pluripotent cells, this histone variant is refractory to the compaction associated with polycomb repression [238]. Interestingly, the polycomb repressed scenario from EC4 to C2 mainly corresponds to MRT changes from late to early replicating loci (Fig. V.12B).

The second mechanism corresponds to transitions from the silent unmarked (EC3) and H2AZ rich accessible (EC4) states to the HP1-associated heterochromatin state C4 (Fig. V.6A). This mechanism corresponds to a dramatic redistribution of the histone modification H3K9me3 which, although present in the pluripotent EC4 state, expands into large (from several 100 kb to a few Mb) late replicating highly compacted heterochromatin (Table V.4, Figs V.7 and V.9). H3K9me3 is important for the formation of the constitutive heterochromatin via the anchoring of the α and β isoforms of the HP1 protein [273–275]. There is also evidence of some crosstalk between H3K9 methyltransferase (HKMT) and DNA methyltransferase (DNMT) [276–280] that might explain the correlation observed between H3K9me3 and DNA methylation and the contribution of the later to the long-term maintenance of these large domains of late replicating C4 heterochromatin devoid of H2AZ and of any other histone modification but H3K9me3 [224, 228]. Importantly the accumulation of such highly condensed, transcriptionally inactive heterochromatin regions

comes along with some subnuclear repositioning towards the nuclear periphery accompanied by important 3D architectural rearrangements [34, 155, 236, 251]. Knockout studies of H3K9 methyltransferases and H3K27 methyltransferases have led to differentiation or development defects [281–286], confirming that the epigenetic mechanisms underlying heterochromatin expansion play a critical role in cell fate determination.

V.3.3 Master replication origins at U/N-domain borders are determinants of cell-fate commitment

We found that MRT changes induced by differentiation resulted in an important change in the number and size of replication U/N-domains [94]. Small neighboring U/N-domains merged to become one large coordinately replicated domains (2 and 3 domains merged to 1 in Figs V.9 left and right column respectively). This replication domain consolidation [34, 35] is thus the consequence of an active early replication initiation zone in ESCs that no longer fires early and is likely replicated passively by a replication wave originating from nearby master replication origins. To characterize this consolidation phenomenon from pluripotent to differentiated cell lines as well as between differentiated cell lines, we defined an index of conservation (Sect. V.5.15) that quantifies the number of U-domain borders in a given cell line that were also shared by (n-1) other cell lines. To the sets of U-domains of the cell types considered so far, namely H1hesc, K562, Monocd14ro1746, Gm12878 and Nhdfad, we added those previously identified in HeLa cells [13, 94] and the germline replication skew N-domains [94, 101–106] (Fig. V.9). For each cell type, about half U-domains are shared by at least another cell line, namely H1hesc (38.4%), K562 (61%), Gm12878 (59.2%), Nhdfad (51.6%) and Ndom (50.2%). Note that the smallest matching percentage was obtained for H1hesc as a direct consequence of the largest number of U-domains in this ESC lines. When looking at U-domain borders individually (peaks in replication timing [13]), we got the following percentages of matching with at least another U/N-domain borders in another cell line: H1hesc (78.8%), K562 (88.1%), Gm12878 (88.9%), Nhdfad (85.6%) and Ndom (87.9%). As originally revealed in skew N-domains [104, 114] and further confirmed in MRT U-domains [94], there exists a remarkable gene organization inside these replication domains that turns out to be robust in each cell type. Expressed genes are confined in the euchromatin C1 (resp. EC1) environment of the bordering master replication origins whereas non expressed

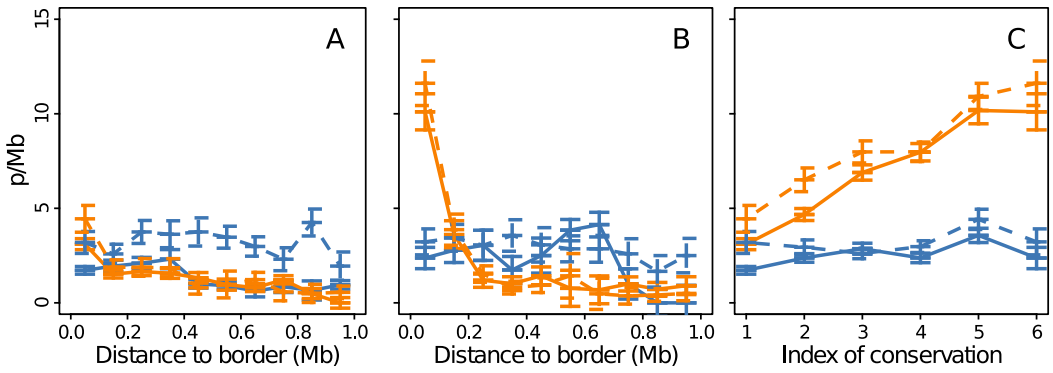

Figure V.15: Distribution of expressed (orange) and not expressed (blue) gene promoters inside replication timing U-domains of H1hesc (solid line) and Nhdfad (dashed line). (A) Mean density of gene promoters with respect to the distance to the closest U-domain border specific to the cell line (n=1). (B) Mean density of gene promoters with respect to the distance to the closest U-domain common to all cell lines (n=6). (C) Mean density of gene promoters in the 100 kb windows containing a U-domain border versus its conservation index n (Sect. V.5.15).

	Nb	GC	CpGo/e	H2AZ	CTCF	NANOG	OCT4	DHS	Promoters /border	Expressed promoters/border
1 to 1	38	0.445	0.233	0.331	0.663	1.321	0.335	0.016	3.947	3.237
1 to 2	3	0.458	0.235	0.346	0.472	0.733	0.145	0.018	3.333	3
2 to 1	4	0.444	0.198	1.214	1.066	1.254	0.133	0.025	3	1.75
2 to 2	6	0.427	0.206	0.606	0.872	0.556	0.458	0.021	2.667	2
2 to 3	1	0.436	0.179	0.468	0.248	0.277	0	0.014	2	2
3 to 2	1	0.418	0.16	0.765	0.419	0	0	0.008	3	3
4 to 2	1	0.391	0.143	0.526	0.261	0	0	0	2	2
4 to 3	1	0.445	0.192	1.227	0.049	0.234	0	0.004	1	1

Table V.7: Sequence, epigenetic and gene characteristics of conserved (n=6) replication U-domain borders of H1hesc that switch from state ECi to Cj.

genes are distributed rather uniformly inside these domains (Fig. V.15) independently of the gradient of chromatin states (Fig. V.13A')(resp. Fig. V.13A). When comparing the gene content nearby replication U/N-borders for increasing index of conservation (Fig. V.15C), we found that the density as well as the distribution of non-expressed genes were quite insensitive to the degree of ubiquitiness of the nearby master replication origin. In other words, non-expressed genes seem to have no knowledge of the replication wave initiating at U/N-domain borders. We got the opposite for expressed genes with a significant enhancement of gene density when increasing the conservation index n (Fig. V.15C). Ubiquitous master replication origins are surrounded by a

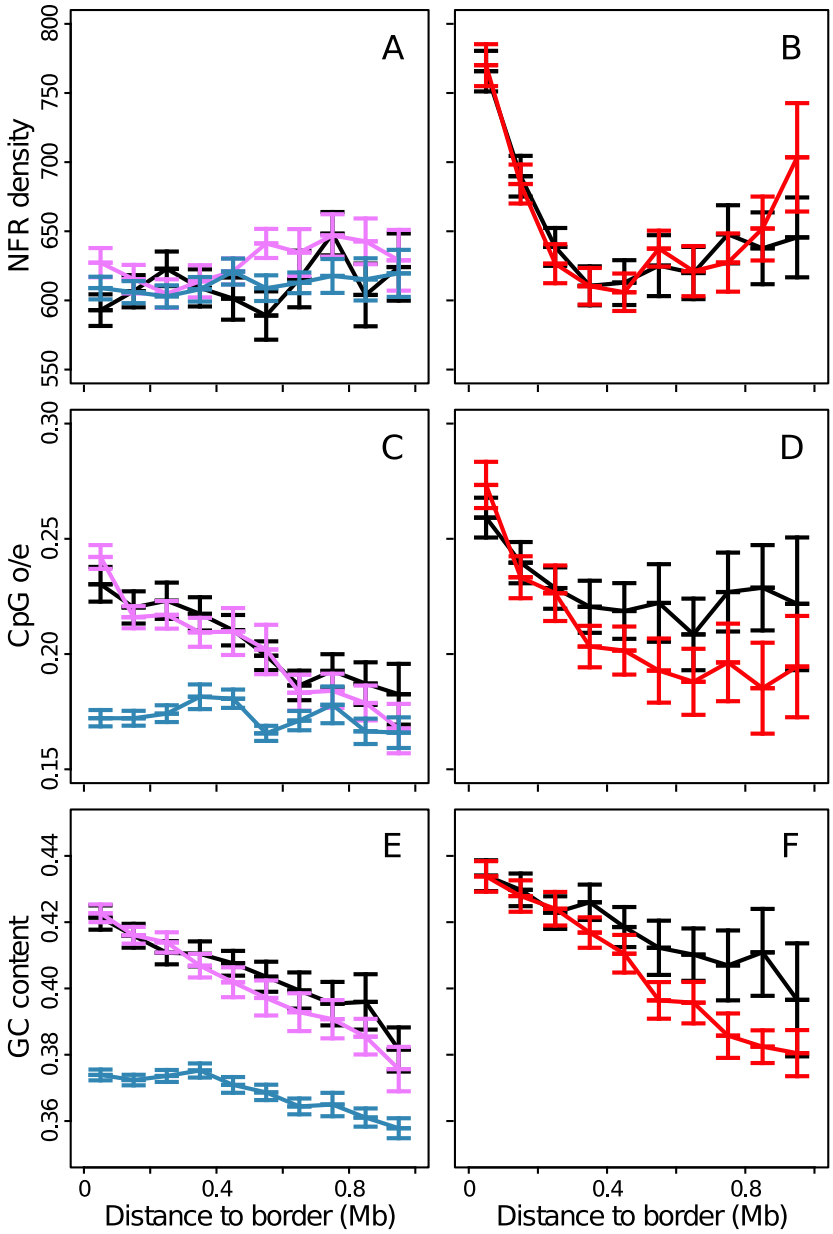

Figure V.16: Sequence characterictic of MRT U-domains of H1hesc and Nhdfad. (A) Density of nucleosome free regions (NFRs) with respect to the distance to the closest U-domains border specific (n=1) to the cell line. The different colors correspond to specific U-domains of Nhdfad (black), specific U-domains of H1hesc whose border is in EC1 or EC2 (red) and specific U-domains of H1hesc whose border is in EC4 (blue). (B) Density of (NFRs) with respect to the distance to the closest conserved (n=6) U-domains border. (C) same as (A) for the CpG o/e. (D) same as (B) for the CpG o/e. (E) same as (A) for the GC content. (F) same as (B) for the GC content.

C1 euchromatin environment which is hypomethylated (Fig. V.16D), GC-high (Fig. V.16F, Table V.7), significantly enriched in DHS and CTCF (Table V.7) and more importantly in nucleosome free regions (NFRs)(Fig. V.16B) coded in the DNA sequence via high energy barriers that impair nucleosome formation (Sect. V.5.12) [112, 203, 287–290]. Thus these ubiquitous master replication origins are specified by an open chromatin structure which is to some extend encoded in the DNA sequence [4, 112]. This also provides some understanding of the local clustering of highly expressed genes with strong CpG rich promoters including house-keeping genes (Fig. V.15B). As exemplified with the Nhdfad cell line, master replication origins that are specific to a differentiated cell line are still GC high (Fig. V.16E) but no longer enriched in NFRs (Fig. V.16A) suggesting that these early firing regions are epigenetically regulated and no longer favored by the DNA sequence. Indeed, Nhdfad specific master replication origins are hypomethylated (Fig. V.16C), and significantly enriched in H2AZ (Fig. V.14A) and CTCF (Fig. V.14B) epigenetic marks. They are mainly surrounded by tissue specific genes with weak CpG poor promoters. Our results are consistent with previous reports that most genes do not change expression during domain consolidation in the MRT profile [12,34,35,88] (Fig. V.8C). Among the expressed genes located in the vicinity of an early master replication origin in a given cell line that experiences a domain consolidation in another cell line, only a small number of genes mainly with CpG poor promoter are repressed whereas CpG rich promoters have the ability to overcome (C4) heterochromatin repression [12, 34, 35]. This coordinated changes in MRT and chromatin state without being accompanied by a global change of the expression program suggests that phenotypic differences between cell types are better reflected by epigenetic properties including the MRT than by transcriptional differences. In that respect, the master replication origins at U/N-domains are likely to be a milestone in the understanding of cell-fate commitment.

V.3.4 ESC specific master replication origins as the cornerstone of pluripotency maintenance

Master replication origins that are specific to the pluripotent H1hesc cell line actually correspond to lineage-independent switches in MRT that are stably maintained after the late epiblast stage. These pluripotent master replication origins (N=483) are almost equally distributed in the chromatin states EC1

	Nb	GC	CpGo/e	H2AZ	CTCF	NANOG	OCT4	DHS	Promoters /border	Expressed promoters/border
1 to 1	58	0.442	0.226	0.361	0.602	1.129	0.159	0.014	3.379	2.655
1 to 2	15	0.429	0.21	0.44	0.655	2.689	0.759	0.013	1.667	1.4
1 to 3	12	0.42	0.201	0.241	0.488	0.998	0.199	0.013	1.917	1.25
1 to 4	14	0.399	0.194	0.469	0.212	0.783	0.13	0.009	0.643	0.643
2 to 1	9	0.418	0.224	0.672	0.886	1.474	0.076	0.019	1.889	1
2 to 2	56	0.441	0.205	0.804	0.903	1.507	0.355	0.017	2.357	1.25
2 to 3	16	0.43	0.181	0.582	0.482	2.418	0.691	0.014	0.875	0.5
2 to 4	19	0.408	0.214	0.977	0.503	2.968	0.76	0.012	0.684	0.158
3 to 2	5	0.374	0.15	0.276	0.336	0.074	0	0.003	1.4	0.6
3 to 3	11	0.371	0.143	0.228	0.069	0.257	0	0.003	0.273	0.273
3 to 4	14	0.352	0.165	0.352	0.068	0.217	0.165	0.004	0.214	0.143
4 to 1	1	0.399	0.231	0.189	0.288	0	0.389	0.008	0	0
4 to 2	13	0.391	0.155	0.824	0.314	0.244	0.148	0.008	1	0.308
4 to 3	22	0.388	0.164	1.007	0.226	0.577	0.257	0.006	0.864	0.409
4 to 4	46	0.376	0.172	0.984	0.14	1.232	0.408	0.006	0.174	0.087
4 to 4 intergenic	39	0.374	0.17	1.013	0.145	1.402	0.481	0.006	0	0

Table V.8: Sequence, epigenetic and gene characteristics of specific (n=1) replication U-domain borders of H1hesc that switch from state ECi to Cj.

(N=113), EC2 (N=131) and EC4 (N=149) and only few are in the unmarked state EC3 (N=51) and in the discarded set D (N=41) (Table V.8). Those that are gene rich in EC1 and EC2 environments display very similar properties than master replication origins specific to differentiated cell lines. They are hypomethylated (Fig. V.16C), enriched in CTCF (Fig. V.14B) and DHS (Table V.7), their GC content is high (Fig. V.16E) but they are not enriched in constitutive NFRs (Fig. V.16A) as an indication of epigenetic regulation. Note that these master replication origins are highly covered by the H2AZ mark but they are nonetheless depleted compared to the very high level coverage of the genome (Fig. V.14A). Somatic specific master replication origins have the same coverage than specific ESC ones, but in contrast they are enriched compared to the genome background (Fig. V.14A). These specific EC1, EC2 master replication origins indeed correspond to the borders of small replication U-domains that overall are replicated in the first half of S-phase. Thus, the associated domain consolidation unlikely involves a switch to late replication as well as important global change in 3D chromatin organization. Consistently these H1hesc specific (EC1, EC2) master replication origins are enriched in the key pluripotency transcription factors NANOG (Fig. V.14C) and OCT4 (Fig. V.14D) (Table V.8).

More surprising is the non negligible proportion (30.9%) of specific H1hesc origins that belong to a EC4 environment and that mainly consolidate into a C4 heterochromatin domain (Table V.8). These EC4 master replication origins indeed correspond to the early replicating EC4 regions that experience a EtoL transition mostly towards the HP1-associated heterochromatin state C4 (Tables V.2 and V.8). They have totally different epigenetic and sequence properties. They are methylated (Fig. V.16C), no longer enriched in CTCF (Fig. V.14B) and DHS (Table V.8), their GC content is low (Fig. V.16E) and they are still not enriched in constitutive NFRs (Fig. V.16A). Actually they are mainly epigenetically regulated by a local enrichment of H2AZ (Fig. V.14A) that turns out to play an unexpected specific role in regulating the spatio-temporal replication program in pluripotent cells. Notably, these pluripotent specific EC4 master replication initiation zones are gene deserts: only 20/82 ($\sim$ 24%) contain a gene promoter as compared to 82/99 (resp. 75/100) for those in EC1 (resp. EC2). Nevertheless, they are enriched in NANOG and OCT4 (Fig. V.14C, D, Table V.8), even in the intergenic ones, which suggests that these transcription factors are also involved in the regulation of replication in pluripotent cells. Note that the unusual principle of chromatin folding

during development reported in [39,40] likely results from the C4 domain consolidation of these early ESC specific EC4 master replication origins (see for example one of them at position 47 Mb on the right panels of Fig. V.9). As discussed in previous works [34–36,39,40], the EtoL transitions associated with the consolidation of pluripotent specific EC1 (see for example one of them at position 12.5 Mb on the left panel of Fig. V.9), EC2 and EC4 to HP1-associated C4 heterochromatin likely coincide with the emergence of compact chromatin near the nuclear periphery and with a dramatic large-scale 3D genome reorganization that may constitute an epigenetic barrier to cellular reprogramming. In that respect, the master-replication origins bordering ESC specific replication U/N-domains are likely to be major determinants in the maintenance of pluripotency.

V.4 Conclusion/Perspectives

In summary, the integrative analysis of genome-wide epigenetic marks, expression and MRT data in an ESCs an differentiated human cell lines, shows that the combinatorial complexity of these epigenetic data can be significantly reduced consistently with previous studies in *Drosophila* [50, 53], *Arabidopsis* [51] and *human* [154, 193, 238]. The epigenetic landscapes of pluripotent and differentiated cells are drastically different even though, in both cases, four but distinct prevalent chromatin states are enough to characterize the diversity in chromatin environment along human chromosomes. Among these four states, only one is transcriptionally active and three are silent. The first one is a gene rich euchromatin state that is shared by pluripotent (EC1) and differentiated (C1) cells as well as the "unmarked" states EC3 and C3 that correspond to a silent state not enriched in any available epigenetic marks. The two other states are different as the signature of the global accessible character of the pluripotent chromatin [236]: H2AZ and H3K4me1 marks are broadly distributed [238] in the bivalent state EC2 containing bivalent genes and in the gene-poor accessible EC4 state as compared to the polycomb repressed state C2 and the HP1-associated heterochromatin state C4 that respectively result from the spreading of H3K27me3 and H3K9me3 in differentiated cells [228,238]. When looking at the way these chromatin states are distributed along human chromosomes with a special focus on the regions where the MRT changes significantly during differentiation, we show that the master replica-

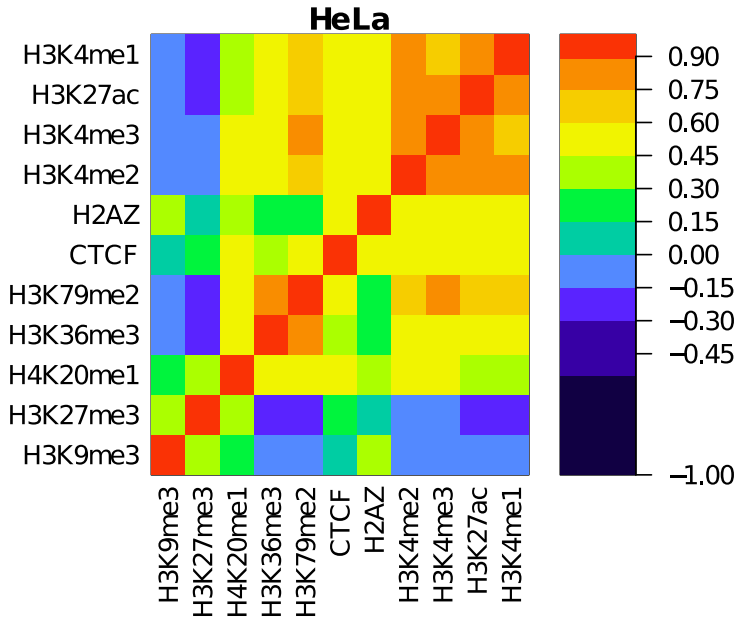

Figure V.17: Spearman correlation matrix between epigenetic marks in HeLa.

tion origins that border megabase-sized MRT U/N-domains [94, 101, 110] are major determinants in cell-fate commitment and lineage fidelity. The minority (5.3%) that are conserved in all cell lines have a peculiar high GC hypomethylated (EC1, C1) euchromatin environment highly enriched in open marks including H2AZ, CTCF, DNase HS and also in NFRs encoded in the DNA sequence suggesting that these ubiquitous master replication origins have been selected during evolution. In these particularly highly decondensed regions are also found numerous CpG rich promoters of highly expressed genes including house-keeping genes. Most of the master replication origins that are cell type specific or shared by a few cell types, still correspond to GC-rich euchromatin mainly regulated epigenetically and no longer favored by a local abundance of NFRs encoded in the DNA sequence. They are mainly surrounded by highly expressed tissue-specific genes. A majority of master replication origins specific to ESCs have rather similar epigenetic properties with a high density of neighboring genes that are regulated by the pluripotency factors NANOG/SOX2(data not available)/OCT4. But what our study has revealed is the existence of a class of ESC specific master replication origins that fire early in a GC-low, gene desert EC4 environment that experiences a change to a compact HP1-associated C4 heterochromatin environment during

differentiation. These master origins have a specific epigenetic regulation that sheds a new light on the unexpected role of both H2AZ and the transcription factors NANOG/SOX2/OCT4 in the maintenance of the replication spatio-temporal program in pluripotent cells. An important proportion (67.4%) of the ESC specific master replication origins indeed correspond to EtoL transitions likely associated with some repositioning towards the nuclear periphery and some large-scale 3D chromatin rearrangements that may hinder cell reprogramming [34–36, 39, 40]. As reported in previous studies of 4C [30] and Hi-C [94, 113] data in differentiated cell lines, these master replication zones at MRT U/N-domain borders act on the one hand as insulators that delimit topological domains of self-interacting chromatin [29, 94], and on the other hand as long-distance interconnected hubs in the intra- and inter- chromosome interaction network [113, 260]. As similar comparative analysis of Hi-C data in ESCs is under progress. Besides confirming the key role played by ESC specific master replication origins in the 3D chromatin regulation and control of pluripotency, we hope to bring new elements of discussion concerning the hypothetized influence of longer G1-phase enabling targeting of loci to the nuclear periphery, and providing more time for nuclei to reorganize their genome before replication initiates in differentiated cells [34,35]. In that respect focusing our study to cancer cells looks very promising (see the particular chromatin structure of HeLa cells revealed on the epigenetic mark matrices presented in Fig. V.17). Recent Meta-analysis [219,291] of replication timing profiles, Hi-C data and somatic copy-number alterations (SCNA) observed in cancer samples from diverse cancer types [292] showed that SCNAs tends to fuse genomic regions that, prior to the rearrangement, spatially co-localized within the nucleus and have similar replication timing. As fundamental structural and functional units underlying the plasticity of replication domain organization in relation to gene expression and chromatin states, the replication timing U/N-domains together with the bordering master replication origins provide a framework for further studies in different cell types and different organisms, in both health and disease.

V.5 Materials and methods

V.5.1 Mean replication timing data and replication U-domain coordinates

Timing profiles for an ESC line (BG02), a lymphoblastoid cell line (GM06990), a skin fibroblast cell line (BJ), an immature myeloid cell line K562 and a HeLa cell line were obtained from the authors [94]. The mean replication timing (MRT) is given for 100 kb non-overlapping windows in hg18 coordinates. We also retrieved the coordinates of the 1534 (BG02), 882 (GM06990), 1150 (BJ), 876 (K562) and 1498 (HeLa) U-domains for the same cell lines from the authors [94].

V.5.2 Histone marks, H2AZ, CTCF, CHD1, NANOG and OCT4 ChIP-seq data

ChIP-seq data were retrieved for the following cell lines: an ESC line (H1hesc), an immature myeloid cell line (K562), a monocytes-CD14+ (monocd14ro1746), a lymphoblastoid cell line (Gm12878), a mammary epithelial cell line (Hmec), an adult dermal fibroblast cell line (Nhdfad).

For all ChIP-seq data, we downloaded data in the Encode standard format "broadpeaks" (`http://genome.ucsc.edu/FAQ/FAQformat.html`). Broadpeaks format is a table of significantly enriched genomic intervals. The score (fold enrichment compared to a uniform distribution of reads) associated with each enriched interval, is the mean signal value across the interval [48, 49]. Most of the data correspond to the release 3 (August 2012) of the Broad histone track. We downloaded the tables from:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/`

The NANOG and OCT4 data corresponds to the release 3 (September 2012) of the HAIB TFBS track. Tables were downloaded from the UCSC from:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/`

For all cell types, we downloaded the broadpeak tables for the following antibodies: CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, H2AZ, H3K79me2, H4K20me1. For the H1hesc cell line, we downloaded these additional broadpeak genomic intervals: CHD1, EZH2, NANOG and OCT4.

V.5.3 Epigenetic profile computation at 100 kb resolution

For each ChIP-seq data and each cell line, we computed a profile at the 100 kb resolution for the 28465 non-overlapping windows corresponding to the sequenced part of the genome. For an antibody, the score in a 100kb window was computed as the sum of the coverage of each significantly enriched interval multiplied by its score.

V.5.4 Treatment of H1hesc data set

We took into account the specificity of H1hesc cell line epigenetic by applying the clustering pipeline described in Chapter III [154] apart from other cell lines. The number of clusters was set to four because it led to the most qualitatively different chromatin states.

V.5.5 Construction of a shared epigenetic space for differentiated cell lines

For the five differentiated cell lines (K562, Monocd14ro1746, Gm12878, Hmec and Nhdfad), we constructed a shared epigenetic space. Once epigenetic profiles at 100kb were computed for each cell line, we concatenated profiles of the same mark together to obtain one vector of $5 \times 28405 = 170970$ windows per mark. Each vector at the 100 kb resolution were transformed with the R function *rank* with option *ties.method=max*.

V.5.6 Rank transformation and Spearman correlation matrix

All statistical computations were performed using the R software (`http://www.r-project.org/`).

In order to compute the Spearman correlation matrix, the epigenetic profiles at 100 kb resolution were transformed with the R function *rank* with option *ties.method=max*. Then we computed the Pearson correlation matrix on the transformed data set. To reorder the matrix in (Figs V.1, and V.2), we computed the Spearman correlation distance $dSCor$ as:

$$dSCor(X, Y) = 1 - SCor(X, Y), \tag{V.1}$$

where $SCor$ is the Spearman correlation. Then, a dendrogram was computed using the R function *hclust* with option *method=average* and with $dSCor$ as dissimilarity.

V.5.7 Principal component analysis

Principal component analysis was performed on the rank transformed dataset using the function *dudi.pca* from the R package *ade4* (see `http://pbil.univ-lyon1.fr/ADE-4` and [167]) with the option *scale=TRUE* (*i.e.* each variable was centered and normalized before the PCA computation). The first four components were retained which accounts for 86% of the dataset variance. We chose four components because the percentage of variance explained drops sharply after the fourth eigenvalue (Fig. V.3A,A'). Clustering was performed in this 4D space.

V.5.8 Clustering strategy

We used Clara algorithm [153] which is an optimization of k-means for large data set. We used the *clara* function implemented in the R package *cluster*. The options were set to: *stand=FALSE, sampsize=500, samples=20, metric=euclidean*.

For the shared differentiated cells, the number of clusters was set to the number of prevalent chromatin states detected in Chapter III [154]. Previously to the merging of dataset into one shared epigenetic space, we checked that,

when applied to each cell individually, the analysis pipeline led to qualitatively the same epigenetic states.

Poorly clustered data points were removed from the set of chromatin states. The silhouette value [168] is a way to quantify how well a point is clustered.

Definition 2 *Given a particular clustering, $C_1, C_2, \ldots, C_k$, of the data in k clusters, let i be a data point and $d(i, C_j)$ the average distance of the data point i to the members of the cluster C_j. Let i be a member of cluster C_l and*

$$a_i = d(i, C_l), \quad b_i = min_{j \neq l}(d(i, C_j)). \tag{V.2}$$

The silhouette value of the data point i is defined as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}. \tag{V.3}$$

A silhouette value below 0 means that the data point is actually closer in average to the points from another cluster than to the ones it has been assigned to. Points with a negative silhouette value are border line allocations. We decided to remove those points from the set of identified chromatin states. Hence chromatin states are groups (clusters) with homogeneous epigenetic features. 91% (resp. 94%) of all 100 kb non-overlapping windows of the human genome were assigned to one of the four chromatin states C1, C2, C3 or C4 (resp. EC1, EC2, EC3 and EC4) in the differentiated (resp. H1hesc) cell lines.

V.5.9 DNase Hypersentive site data

DNaseI hypersensitive sites (DHSs) data were downloaded in the Encode standard format "narrowpeaks" (`http://genome.ucsc.edu/FAQ/FAQformat.html`). DHS narrowpeaks are genomic intervals indentified as hypersentive zones to DNaseI within a FDR of 0.5%. We downloaded the tables from:

`http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/`

V.5.10 Annotation and Expression data

As human gene coordinates, we used the UCSC Known Genes table. When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates correspond to the union of all the overlapping gene coordinates, resulting in 23818 distinct genes.

Expression data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC). To construct our expression data set, we used RefSeq Genes track as human gene coordinates. Genes with alternative splicing were merged into one transcript by taking the union of exons. Hence the transcription start site (TSS) was placed at the beginning of the first exon. We obtained a table of 23329 genes. We downloaded expression values from the release 2 of Caltech RNA-seq track (ENCODE project at UCSC):

`http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeCaltechRnaSeq/`

Expression for one transcript is given in reads per kilobase of exon model per million mapped reads (RPKM) [10]. RPKM is defined as:

$$R = \frac{10^9 C}{NL}, \tag{V.4}$$

where C is the number of mappable reads that fall into gene exons (union of exons for genes with alternative splicing), N is the total number of mappable reads in the experiment, and L is the total length of the exons in base pairs. We associated 17872 genes with a valid RPKM value in K562 and Gm12878 and 17463 in H1hesc.

V.5.11 CpG o/e computation and GC content

CpG observed/expected ratio (CpG o/e) was computed as $\frac{n_{CpG}}{L-l} \times \frac{L^2}{n_C n_G}$, where n_C, n_G and n_{CpG} are the numbers of C, G and dinucleotides CG, respectively, counted along the sequence, L is the number of nonmasked nucleotides and l is the number of masked nucleotide gaps plus one, *i.e.* L-l is the number of dinucleotide sites. The CpG o/e was computed over the sequence after masking annotated CGIs. The GC content was computed on the native sequence.

V.5.12 Nucleosome free regions (NFR)

The coordinates of the NFRs predicted by the physical model defined in [203, 287–289] were obtained from the authors [290]. This theoretical model amounts to compute the energy required for nucleosome formation based on sequence-dependent bending properties [4].

V.5.13 Chromatin state blocks

We detected contiguous windows of the same chromatin state (C1 to C4 and EC1 to EC4). We then kept the coordinates of the blocks of contiguous windows. To form chromatin state blocks of states (1+2), we merely detected contiguous windows of state 1 or 2. The same procedure was applied to define chromatin blocks of states (3+4). For chromatin blocks (1+2) and (3+4), we authorized the inclusion of isolated windows which did not belong to any chromatin state so to not disrupt very long blocks.

V.5.14 Replication N-domains

The coordinates of the 678 human replication N-domains for assembly hg17 were obtained from the authors [104] and mapped using LiftOver to hg18 coordinates; we kept only the 663 N-domains that had the same size after conversion [94].

V.5.15 Index of conservation for U-domain borders

To identify MRT U-domain borders which are common to several cell lines, we constructed a counting signal and we attributed a conservation index as follows:

1. We created a merged data set of the coordinates of all U-domain borders detected in [94] and of skew N-domain borders detected in [104]. U-domains where detected in the following cell line (BG02, K562, GM06990, H0287, TL010, BJ, HeLa). GM06990, H0287, TL010 are three lymphoblastoid cell lines. To avoid lymphoblastoid cell specific U-domains getting an artificially high conservation index, we took only GM06690

into account. To avoid MRT to be a confounding factor, we excluded late U-domain borders with MRT > 0.5.

2. Then, we slided a 200 kb window along the genome with 10 kb incremental steps. At each position, we retrieved the number of cell lines that have a domain border in the window. By doing so, we constructed the counting signal called the conservation index. For instance, if a U-domain border of K562 has a conservation index of 3, it means that 2 domain borders from other cell lines are at maximum distance of 200 kb.

Chapter VI

General discussion

> The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which ofttimes they are unable to account.
>
> Laplace

This thesis provides a natural way to handle challenges generated by high-throughput sequencing. As discussed in Chapter I, data availability has profoundly changed the way biology is addressed; it is now possible to produce new results by simple analysis of the existing raw data or by conducting a meta-analysis of data produced by different teams. Hopefully, the meta-analyses of existing CHiP-seq, gene expression and MRT data presented in this thesis will help elucidating nuclear functions in human cells.

In this concluding chapter we would like to briefly summarize the results obtained in this thesis articulating them into a speculative model of chromatin structure and of how replication proceeds through it. For a detailed discussion of results chapter by chapter, the reader should refer to the discussions in Sects. III.3, IV.5 and V.4. We discuss what should guide the choice of statistical methods to conduct similar integrative analyses. We also discuss the interpretation of statistical analyses and their use in nowadays biology. We end with ongoing work and perspectives.

VI.1 Summary of results

In this thesis, we applied a simple data workflow based on PCA+clustering (Chapter II) to analyze a total of $\sim$ 100 CHiP-seq profiles in diverse cases.

In chapter III, we demonstrated that the human chromatin in the immature myeloid K562 cell line could be described by four chromatin states with markedly different epigenetic contents. These chromatin states are namely: C1, the state that contains all active transcription marks (*e.g.* H3K36me3, H3K4me3, H3K79me3, etc), C2 characterized by the mark H3K27me3 as an indication of polycomb complex repression, C3 an unmarked state and C4 enriched in the mark H3K9me3 as the signature of the presence of HP1 protein. Our analysis aimed at characterizing the large-scale chromatin structure and its relationship to replication. Therefore, the analyses were conducted at the resolution of MRT, namely 100 kb. These chromatin states were shown to have different MRT distributions and their ordering according to their MRT is C1, C2, C3, C4. We confirmed the well known observation that euchromatin (C1) is replicated early whereas heterochromatin (C3, C4) is replicated late. We established that the polycomb repressed state (C2) is replicated in mid-S phase and not at the end of S phase as its transcriptionally silent status would have predicted. Schematically, C1 contains almost all active promoters but also a lot of inactive genes, C2 contains inactive genes whereas C3 and C4 are gene depleted. In terms of sequence, C1 and C2 are GC-rich and CpG rich whereas C3 and C4 are GC-poor and CpG-poor. The spatial organization of these chromatin states is striking in MRT U-domains. The early replication zone bordering U-domains are mostly covered by the active transcription chromatin state C1. Further from the border, the majority of loci are covered by the chromatin state C2. Then, replication ends in C3+C4 regions.

In chapter IV, we applied our integrative analysis to promoters. We locally (in 6 kb windows) analyzed the structure of chromatin around promoters in the immature myeloid cell line K562. We also found that promoters could be classified in 4 epigenetic classes corresponding to the chromatin states found genome wide at 100 kb resolution namely P1 transcriptionally active, P2 polycomb repressed, P3 unmarked and P4 HP1 repressed. This epigenetic classification partially matches the classification in CpG-rich and CpG-poor promoters reported in [114, 136]. The vast majority of P1 promoters are CpG-rich whereas there are as many CpG-rich promoters as CpG-poor promoters in silent states P2, P3 and P4. Thus, we did not witness a major discrep-

ancy in the epigenetic regulation of CpG-rich and CpG-poor promoters. For instance, one silencing pathway could have been specifically for CpG-poor promoters. Alternatively, we could have identified two active epigenetic classes, one corresponding; to CpG-rich promoters and the other one to CpG-poor promoters. Actually, differences have been identified between the chromatin structures of CpG-rich and CpG-poor promoters. The differences do not reside in the epigenetic marks contents but rather in the nucleosome structure and the spatial distribution of marks on nucleosomes [293, 294]. To address these questions properly, one needs to look at marks at the nucleosome resolution and also to take into account new data on DNA methylation. Even though the resolution of our analysis may be too coarse for the foregoing questions, we successfully uncovered an unexpected interplay between the small-scale (6 kb) and the larger-scale (100 kb) classifications. In particular, we showed that only one active promoter is enough to imply an active large-scale environment. This "specification" is not affected by the number of inactive genes surrounding the active gene. On the contrary, a large heterochromatin (C4) environment implies that genes are inactive (belonging to P3 and P4). Therefore active genes seem to determine large scale genome organization whereas the large scale classification of inactive genes is subjected to their environment (presence of active genes or spreading of heterochromatin). We also revealed that promoters exhibit a striking organization inside U-domains with highly expressed genes confined in the burst ($\sim$ 200 kb) of open chromatin at U-domain borders confirming previous results about the organization of transcription in their germline counterpart [114, 136].

In Chapter V, we extended our analysis to several cell lines including one embryonic stem cell (ESC). As a first general result, we showed that the four chromatin states described in Chapter III are general to all somatic cell lines. However, the MRT ordering of chromatin states C3 and C4 seems to be cell line dependent. As a major result, our analysis revealed in ESC, a chromatin structure that departs from the general organization observed in somatic cells. Indeed, ESCs have drastically different chromatin states: EC1 is transcriptionally active, EC2 is a bivalent state (*i.e.* containing both active promoter H3K4me3 and H3K27me3), EC3 is unmarked and EC4 is a silent dynamic chromatin. We further showed that U-domain borders conserved in all cell lines have a particular sequence and epigenetic signature. Indeed, conserved U-domain borders have a particularly high active chromatin state coverage and they are GC and CpG rich and significantly enriched in intrinsic (encoded

in the DNA sequence) NFRs. At the other end of the spectrum, specific U-domain borders differ in ESC and sommatic cell lines. In the later, they are marked by a C1+C2 open chromatin enrichment and a noticeable absence of heterochromatin C3+C4 environment, probably corresponding to tissue specific transcription zones. In ESC, specific borders can be divided in two subclasses. The first behaves like specific borders in somatic cell line. The second class is a particularity of pluripotent cells with early initiation zones occurring in gene deserts, GC poor EC4 regions. This peculiarity is enabled by the dynamic features of chromatin state EC4 (as the presence of chromatin remodelers). These early replicating zones do not seem to have any signature in their sequence and the vast majority of them are intergenic. By contrast they have a clear epigenetic signature since they are enriched in the histone variant H2AZ as well as in the pluripotent factors NANOG and OCT4, putting into light the role of these epigenetic marks in the regulation of the spatio-temporal replication program in ES cells.

VI.2 Putative model for the interplay between chromatin and replication

By taking into account the results reported in this thesis and in previous studies, we can elaborate on a model of replication proceeding through four different chromatin states. Chapter V demonstrates the generality of the chromatin segmentation in four chromatin states. Results from chapter IV suggest that large-scale transcriptionally active regions are specified by the presence of active genes (which does not exclude inactive genes). Replication origins in these large-scale active regions fire early [127,128]. However, in chapter V, the example of ESC specific early replication initiation in gene desert demonstrates clearly that transcription is not necessary for early initiation. Seemingly, chromatin has to be accessible for origins to fire early in the S-phase. In somatic cell lines, the main nuclear function promoting accessibility seems to be transcription. Yet, the example of ESC shows that chromatin accessibility can be obtained by a specific regulation. Then, as suggested in [93], the progression of chromatin forks could open neighboring chromatin and stimulate origins in more compact chromatin. This is consistent with the presence of origins in transition timing regions (TTR) [93,94] and the succession of chromatin states (C1, C2, C3, C4) visited by the accelerating replicating wave inside MRT U-

domains, from the more open at the border to the more compact at the center. When reaching heterochromatin states (C3 or C4), origin positions of origin likely become random as suggested by the increased fork density detected in [93, 109] and the few well positioned ORC complexes observed in [77].

Previous analyses on the 3D structure of the genome allow us to elaborate on the position of early firing regions in the nucleus. In [30], a 4C assay revealed that U-domain borders preferentially interact together and FISH experiments confirmed their spatial proximity in the nucleus. Late replicating regions also preferentially interact with late replicating regions. Yet late U-domain central regions preferentially interact with loci in the same U-domain, U-domain borders acting as barriers. These analyses were further confirmed by quantitative analysis of the graph of interactions generated from the Hi-C matrix in which U-domains border appear as hubs of intra- and inter- chromosome interactions [113]. In [94], U-domains where shown to be enriched in CTCF which forms chromatin loops and act as a barrier that prevent the heterochromatin from spreading. Altogether, these observations suggests that replication starts at the center of the nucleus and propagates towards more peripheral zones segregated from each other. Also, the Hi-C interaction matrix structure in U-domains shows sub-domains of interactions suggesting successive folding levels. The existence of three kinds of silent chromatin states (C2, C3, C4) inside U-domains might imply preferential interaction with chromatin state of the same class. Further analysis of the Hi-C results conditioned by chromatin states and theoretical modeling of the chromatin fiber by taking into account the four types of chromatin states could provide answer to these questions.

Our simple statistical description together with recent work naturally raise a lot of speculations. Proposing speculative models is important to construct interesting hypotheses to work on. However one has to be aware of the difference between correlation and causality to avoid overstating and/or misinterpreting statistical results. Also, to interpret correctly a statistical model, the biostatisticians should know how to proceed to choose models being aware that there is no absolute optimum model.

VI.3 Designing a statistical analysis: a question of choice

> All models are wrong but some models are useful
>
> George Box

An issue with statistical analyses is the choice of the method. It would be very handy to have a method that is objectively superior to others for all types of problems. The "no free lunch" theorem formally states the opposite [295]. To choose the proper machine learning device, the biostatistician has to know what he would like to achieve in order to find a method that is specifically efficient for this particular task. Is it to predict a variable precisely according to other variables without necessarily willing to understand the causal links between those variables? Is it to find homogeneous groups within a set of individuals characterized by several variables?

An important distinction between statistical methods is the classification as supervised or unsupervised methods (see the introduction in [45] on the subject). In the first case, a model is constructed from a set of D explanatory variables X_i to fit a variable of interest Y:

$$Y = f(X_1, X_2, \ldots, X_D) + \varepsilon, \tag{VI.1}$$

where f is a statistical model and ε is the error of the model. The parameters of the function f are optimized so that the error ε of the model is minimal. The advantage is that the prediction error ε can be evaluated on an independent dataset for the same parameters. Therefore, there exists a natural way to assess the quality of a supervised model. The shortcoming of supervised methods is that only features that predict Y are retained in the explanatory variables. In the unsupervised case, the goal is to describe the behavior of a set of variables without distinguishing between explanatory variables and the variable to explain (usually of more interest). The quality of unsupervised methods is more difficult to assess (because there is no formal prediction error). Yet, all particularities of the dataset are explored.

In this thesis, we wanted to describe the chromatin structure independently from replication, and afterwards see how the chromatin structure impacts

replication. This justified the choice of unsupervised methods. An alternative would have been to apply a supervised method (*e.g.* linear regression, knn-classification) to predict replication timing from epigenetic marks. We would probably have reached the same overall conclusions. However, we can wonder if we would have found the distinction between the unmarked chromatin state and HP1 associated heterochromatin. Indeed, the timing ordering of these two chromatin states is cell line dependent. Since the differences between these states is not important in terms of replication timing, a supervised method may have missed this distinction that turns out to be quite relevant from the epigenetic standpoint.

VI.4 Causality and correlation: risk of overstatements

Whatever the strength of the statistical link/correlation between two variables, it does not imply the existence of a causal link between the variables. As clearly exemplified in [296], pear crops are not mechanically linked to apple crops even if the two variables are dependent. Indeed, in a year with favorable weather conditions, both apple crops and pear crops would be good. By contrast, during a harsh year, pear and apple crops would be equally bad. Therefore apple and pear crops correlate. The mechanical link here is the shared weather conditions. The funny fact is that, we could imagine that a model based on apple crops to predict pear crops would be better than a model based on a mechanical factor like precipitation. Indeed, apple crops summarize several weather variables (*e.g.* precipitations, hours of sunshine) in one variable. More than an "egg-and-chicken" problem, the statistical link between transcription and replication may be more like the link between the apples and the pears. The hidden causal variable "weather conditions" would be the chromatin accessibility (*e.g.* local openness, 3D position in the nucleus). However, in contrast to apples and pears that do not affect the weather, replication and transcription machineries are known to directly act (feedback) on chromatin accessibility.

Models based on statistical analyses can only be speculative. [297] elegantly stressed this idea by qualifying genomic data analysis as hypothesis-generating biology. All these hypotheses have to be verified by what is referred to as the "hypothesis-based" biology in [297]. From our standpoint, a classical biol-

ogy assay formulates an hypothesis on causal reasons of an observation and then rigorously tests it, possibly by applying classical hypothesis tests (for a clear exposition of the hypothesis test approach see [298]). Multivariate data analysis provides biology with interesting leads in the space to explore. The hypothesis-generating part is essential because the initial number of possibilities is unmanageable. An even more accurate name for this type of research could be "hypothesis-confining" biology.

We hope that our work will make the study of causal links between chromatin, transcription and replication easier. One direct extension of this work would be to take advantage of the recently avaible genome-wide datasets of origin positions in human to analyze the impact of chromatin structure on origin positioning. There are currently several genome-wide available datasets. In some studies, ORIs were detected using nascent DNA strands [73, 77]. In other studies, ORC were detected by CHiP-seq [128]. Two teams which belong to our collaborators are currently positioning origins genome-wide by sequencing nascent DNA strands [299] and Okazaki fragments [300]. The availability of origin positions in several cell lines will enable us to assess the impact of chromatin context on origin positioning and firing time. Moreover, from the comparison of datasets obtained by different techniques, we will likely ensure the robustness of results. A second very promising extension would be to apply a similar analysis as the one presented in Chapter V to the development of cancer. Our statistical workflow revealed particularities of ESC cell lines and detected profound epigenetic rearrangements in cancer cells (as exemplified on Hela cells in Fig. V.17). However, for specific application to cancer, the copy number variations should be taken into account [301]. Apart from this precaution, systematic comparison between healthy cell lines and their cancerous counterparts using our statistical procedure looks very promising. Our team and our collaborators are currently developping a project aiming at integratively characterizing nuclear rearrengement during MCF-7 cell line proliferation. About 75% of breast cancers rely on estrogen for proliferation. MCF-7 is an estrogen receptor-α (ERα) positive breast cancer cell line that has been successfully used for more than 40 years to study estrogen response in breast cancer [302]. We propose to establish the replication profile for this cell line before and at three time points after estrogen treatment. Given these 4 time points, we will have for the first time the opportunity to analyze changes in the replication timing profile along a well defined temporal path of during the progression of cancer. An important reason why we chose this system is

that it has been extensively analyzed so that a very large amount of data is available for this cell line. For example, the ENCODE project currently lists 116 experiments for MCF7. The application of our statistical procedure at the four time points would enable to go beyond static correlative analysis of the relationships between replication timing, chromatin states, gene expression and chromosome topological structure and to address important issues like whether the epigenetic and structural modifications happen concomitantly or in an ordered manner. It will likely lead to interesting hypotheses to understand the epigenetic deregulation occurring in cancer.

Bibliography

1. Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.

2. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470: 187–197.

3. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.

4. Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafa Y, et al. (2011) Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Phys Rep 498: 45–188.

5. Allis CD, Jenuwein T, Reinberg D, Caparros ML, editors (2006) Epigenetics. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

6. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12: 7–18.

7. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

8. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28: 1045–1048.

9. The ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (encode). PLoS Biol 9: e1001046.

10. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5: 621–628.

11. Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, et al. (2004) Replication timing of the human genome. Hum Mol Genet 13: 191–202.

12. Desprat R, Thierry-Mieg D, Lailler N, Lajugie J, Schildkraut C, et al. (2009) Predictable dynamic program of timing of DNA replication in human cells. Genome Res 19: 2288–2299.

13. Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, et al. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res 20: 447–457.

14. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, et al. (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci USA 107: 139–144.

15. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669–681.

16. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.

17. Ozsolak F, Song JS, Liu XS, Fisher DE (2007) High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol 25: 244–248.

18. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887–898.

19. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, et al. (2011) Determinants of nucleosome organization in primary human cells. Nature 474: 516–520.

20. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3: 511–518.

21. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311–322.

22. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science 295: 1306–1311.

23. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, et al. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38: 1348–1354.

24. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 38: 1341–1347.

25. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, et al. (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16: 1299–1309.

26. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.

27. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462: 58–64.

28. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol 30: 90–98.

29. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.

30. Moindrot B, Audit B, Klous P, Baker A, Thermes C, et al. (2012) 3D chromatin conformation correlates with replication timing and is conserved in resting cells. Nucleic Acids Res 40: 9470–9481.

31. Holwerda S, de Laat W (2012) Chromatin loops, gene positioning, and gene expression. Front Genet 3: 217.

32. Dostie J, Bickmore WA (2012) Chromosome organization in the nucleus–charting new territory across the Hi-Cs. Curr Opin Genet and Dev 22: 125–131.

33. Cavalli G, Misteli T (2013) Functional implications of genome topology. Nat Struct Mol Biol 20: 290–299.

34. Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, et al. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 6: e245.

35. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, et al. (2010) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res 20: 155–169.

36. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, et al. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res 20: 761–770.

37. Zhou J, Ermakova OV, Riblet R, Birshtein BK, Schildkraut CL (2002) Replication and subnuclear location dynamics of the immunoglobulin heavy-chain locus in B-lineage cells. Mol Cell Biol 22: 4876–4889.

38. Williams RR, Azuara V, Perry P, Sauer S, Dvorkina M, et al. (2006) Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. J Cell Sci 119: 132–140.

39. Takebayashi S, Dileep V, Ryba T, Dennis JH, Gilbert DM (2012) Chromatin-interaction compartment switch at developmentally regulated chromosomal domains reveals an unusual principle of chromatin folding. Proc Natl Acad Sci USA 109: 12574–12579.

40. Takebayashi S, Ryba T, Gilbert DM (2012) Developmental control of replication timing defines a new breed of chromosomal domains with a novel mechanism of chromatin unfolding. Nucleus 3: 500–507.

41. Berezney R (2002) Regulating the mammalian genome: the role of nuclear architecture. Adv Enzyme Regul 42: 39–52.

42. Zink D, Bornfleth H, Visser A, Cremer C, Cremer T (1999) Organization of early and late replicating DNA in human chromosome territories. Exp Cell Res 247: 176–188.

43. Cook PR (1999) The organization of replication and transcription. Science 284: 1790–1795.

44. Grasser F, Neusser M, Fiegler H, Thormeyer T, Cremer M, et al. (2008) Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. J Cell Sci 121: 1876–1886.

45. Murphy KP (2012) Machine Learning a probabilistic perspective. Cambridge, Massachusetts: the MIT press.

46. Mayer Schönberger V, Cukier K (2013) Mise en données du monde, le déluge numérique. Le Monde Diplomatique .

47. Marx V (2013) Biology: The big challenges of big data. Nature 498: 255–260.

48. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817–825.

49. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the encode and modencode consortia. Genome Res 22: 1813–1831.

50. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell 143: 212–224.

51. Roudier F, Ahmed I, Bérard C, Sarazin A, Mary-Huard T, et al. (2011) Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J 30: 1928–1938.

52. Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, et al. (2011) Broad chromosomal domains of histone modification patterns in *C.elegans*. Genome Res 21: 227–236.

53. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, et al. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell 148: 458–472.

54. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P (2005) Replication and transcription: shaping the landscape of the genome. Nat Rev Genet 6: 669–677.

55. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693-705.

56. Maric C, Prioleau MN (2010) Interplay between DNA replication and gene expression: a harmonious coexistence. Curr Opin Cell Biol 22: 277–283.

57. Gilbert DM (2010) Evaluating genome-scale approaches to eukaryotic DNA replication. Nat Rev Genet 11: 673–684.

58. Berezney R, Dubey DD, Huberman JA (2000) Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma 108: 471–484.

59. Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. Annu Rev Biochem 71: 333–374.

60. Gilbert DM (2001) Making sense of eukaryotic DNA replication origins. Science 294: 96–100.

61. Méchali M (2010) Eukaryotic DNA replication origins: many choices for appropriate answers. Nat Rev Mol Cell Biol 11: 728–738.

62. MacAlpine DM, Almouzni G (2013) Chromatin and DNA replication. Cold Spring Harb Perspect Biol 5: a010207.

63. Bogan JA, Natale DA, Depamphilis ML (2000) Initiation of eukaryotic DNA replication: conservative or liberal? J Cell Physiol 184: 139–150.

64. Méchali M (2001) DNA replication origins: from sequence specificity to epigenetics. Nat Rev Genet 2: 640–645.

65. McNairn AJ, Gilbert DM (2003) Epigenomic replication: linking epigenetics to DNA replication. Bioessays 25: 647–656.

66. Aladjem MI (2007) Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nat Rev Genet 8: 588–600.

67. Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, et al. (2008) Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature 455: 557–560.

68. Hamlin JL, Mesner LD, Lar O, Torres R, Chodaparambil SV, et al. (2008) A revisionist replicon model for higher eukaryotic genomes. J Cell Biochem 105: 321–329.

69. Costas C, de la Paz Sanchez M, Stroud H, Yu Y, Oliveros JC, et al. (2011) Genome-wide mapping of **Arabidopsis thaliana* origins of DNA replication and their associated epigenetic marks. Nat Struct Mol Biol 18: 395–400.

70. Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, et al. (2011) Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res 21: 1438–1449.

71. Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, et al. (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet 5: e1000446.

72. Lucas I, Palakodeti A, Jiang Y, Young DJ, Jiang N, et al. (2007) High-throughput mapping of origins of replication in human cells. EMBO Rep 8: 770–777.

73. Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, et al. (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci USA 105: 15837–15842.

74. Karnani N, Taylor CM, Malhotra A, Dutta A (2010) Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. Mol Biol Cell 21: 393–404.

75. Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL, et al. (2011) Bubble-chip analysis of human origin distributions demonstrates

on a genomic scale significant clustering into zones and significant association with transcription. Genome Res 21: 377–389.

76. Martin MM, Ryan M, Kim R, Zakas AL, Fu H, et al. (2011) Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res 21: 1822–1832.

77. Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, et al. (2012) Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat Struct Mol Biol 19: 837–844.

78. Hamlin JL, Mesner LD, Dijkwel PA (2010) A winding road to origin discovery. Chromosome Res 18: 45–61.

79. Valenzuela MS, Chen Y, Davis S, Yang F, Walker RL, et al. (2011) Preferential localization of human origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA sequences. PLoS One 6: e17308.

80. Cayrou C, Coulombe P, Puy A, Rialle S, Kaplan N, et al. (2012) New insights into replication origin characteristics in metazoans. Cell Cycle 11: 658-667.

81. Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, et al. (2001) Replication dynamics of the yeast genome. Science 294: 115–121.

82. Lee TJ, Pascuzzi PE, Settlage SB, Shultz RW, Tanurdzic M, et al. (2010) *Arabidopsis thaliana* chromosome 4 replicates in two phases that correlate with chromatin state. PLoS Genet 6: e1000982.

83. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330: 1775–1787.

84. Schübeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, et al. (2002) Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. Nat Genet 32: 438–442.

85. MacAlpine DM, Rodriguez HK, Bell SP (2004) Coordination of replication and transcription along a Drosophila chromosome. Genes Dev 18: 3094–3105.

86. Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C, et al. (2008) Global organization of replication time zones of the mouse genome. Genome Res 18: 1562–1570.

87. Hiratani I, Takebayashi S, Lu J, Gilbert DM (2009) Replication timing and transcriptional control: beyond cause and effect part II. Curr Opin Genet Dev 19: 142–149.

88. Schwaiger M, Stadler MB, Bell O, Kohler H, Oakeley EJ, et al. (2009) Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. Genes Dev 23: 589–601.

89. Farkash-Amar S, Simon I (2010) Genome-wide analysis of the replication program in mammals. Chromosome Res 18: 115–125.

90. Yaffe E, Farkash-Amar S, Polten A, Yakhini Z, Tanay A, et al. (2010) Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet 6: e1001011.

91. Buongiorno-Nardelli M, Micheli G, Carri MT, Marilley M (1982) A relationship between replicon size and supercoiled loop domains in the eukaryotic genome. Nature 298: 100–102.

92. Conti C, Sacca B, Herrick J, Lalou C, Pommier Y, et al. (2007) Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. Mol Biol Cell 18: 3059–3067.

93. Guilbaud G, Rappailles A, Baker A, Chen CL, Arneodo A, et al. (2011) Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. PLoS Comput Biol 7: e1002322.

94. Baker A, Audit B, Chen CL, Moindrot B, Leleu A, et al. (2012) Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. PLoS Comput Biol 8: e1002443.

95. Jackson DA, Pombo A (1998) Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J Cell Biol 140: 1285–1295.

96. Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, et al. (1998) Spatial and temporal dynamics of DNA replication sites in mammalian cells. J Cell Biol 143: 1415–1425.

97. Leonhardt H, Rahn HP, Weinzieri P, Sporbert A, Cremer T, et al. (2000) Dynamics of DNA replication factories in living cells. J Cell Biol 149: 271–280.

98. Cook PR (2001) Principles of Nuclear Structure and Functions. New York: Wiley.

99. Carter DRF, Eskiw C, Cook PR (2008) Transcription factories. Biochem Soc Trans 36: 585–589.

100. Chambeyron S, Bickmore WA (2004) Does looping and clustering in the nucleus regulate gene expression? Curr Opin Cell Biol 16: 256–262.

101. Audit B, Baker A, Chen CL, Rappailles A, Guilbaud G, et al. (2013) Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. Nat Protoc 8: 98–110.

102. Brodie of Brodie EB, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, et al. (2005) From DNA sequence analysis to modeling replication in the human genome. Phys Rev Lett 94: 248103.

103. Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, et al. (2005) Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci USA 102: 9836–9841.

104. Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, et al. (2007) Human gene organization driven by the coordination of replication and transcription. Genome Res 17: 1278–1285.

105. Baker A, Nicolay S, Zaghloul L, d'Aubenton-Carafa Y, Thermes C, et al. (2010) Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes. Appl Comput Harmon Anal 28: 150–170.

106. Chen CL, Duquenne L, Audit B, Guilbaud G, Rappailles A, et al. (2011) Replication-associated mutational asymmetry in the human genome. Mol Biol Evol 28: 2327–2337.

107. Baker A, Julienne H, Chen CL, Audit B, d'Aubenton Carafa Y, et al. (2012) Linking the DNA strand asymmetry to the spatio-temporal replication program. I. About the role of the replication fork polarity in genome evolution. Eur Phys J E 35: 92.

108. Baker A, Chen CL, Julienne H, Audit B, d'Aubenton Carafa Y, et al. (2012) Linking the DNA strand asymmetry to the spatio-temporal replication program: II. Accounting for neighbor-dependent substitution rates. Eur Phys J E 35: 123.

109. Hyrien O, Rappailles A, Guilbaud G, Baker A, Chen CL, et al. (2013) From simple bacterial and archeal replicons to replication N/U-domain. J Mol Biol in press.

110. Audit B, Zaghloul L, Baker A, Arneodo A, Chen CL, et al. (2012) Megabase replication domains along the human genome: relation to chromatin structure and genome organisation. Subcell Biochem 61: 57–80.

111. Audit B, Nicolay S, Huvet M, Touchon M, d'Aubenton Carafa Y, et al. (2007) DNA replication timing data corroborate in silico human replication origin predictions. Phys Rev Lett 99: 248102.

112. Audit B, Zaghloul L, Vaillant C, Chevereau G, d'Aubenton-Carafa Y, et al. (2009) Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells. Nucleic Acids Res 37: 6064–6075.

113. Boulos R, Arneodo A, Jensen P, Audit B (2013) Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. Phys Rev Lett 111: 118102.

114. Zaghloul L, Baker A, Audit B, Arneodo A (2012) Gene organization inside replication domains in mammalian genomes. C R Mécanique 340: 745–757.

115. Ryba T, Hiratani I, Sasaki T, Battaglia D, Kulik M, et al. (2011) Replication timing: a fingerprint for cell identity and pluripotency. PLoS Comput Biol 7: e1002225.

116. Jacob F, Brenner S, Cuzin F (1963) On the regulation of DNA replication in bacteria. Cold Spring Harb Symp Quant Biol 28: 329–342.

117. DePamphilis ML, editor (2006) DNA Replication and Human Disease. Cold Spring Harbor, New-York: Cold Spring Harbor Laboratory Press.

118. Hyrien O, Méchali M (1993) Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. EMBO J 12: 4511–4520.

119. Gerbi SA, Bielinsky AK (2002) DNA replication and chromatin. Curr Opin Genet Dev 12: 243–248.

120. Anglana M, Apiou F, Bensimon A, Debatisse M (2003) Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. Cell 114: 385–394.

121. Fisher D, Méchali M (2003) Vertebrate HoxB gene expression requires DNA replication. EMBO J 22: 3737–3748.

122. Coverley D, Laskey RA (1994) Regulation of eukaryotic DNA replication. Annu Rev Biochem 63: 745–776.

123. Sasaki T, Sawado T, Yamaguchi M, Shinomiya T (1999) Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalpha-dE2F locus of Drosophila melanogaster. Mol Cell Biol 19: 547–555.

124. Gilbert DM (2004) In search of the holy replicator. Nat Rev Mol Cell Biol 5: 848–855.

125. Demeret C, Vassetzky Y, Méchali M (2001) Chromatin remodelling and DNA replication: from nucleosomes to loop domains. Oncogene 20: 3086–3093.

126. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM (2010) Conserved nucleosome positioning defines replication origins. Genes Dev 24: 748–753.

127. Eaton ML, Prinz JA, MacAlpine HK, Tretyakov G, Kharchenko PV, et al. (2011) Chromatin signatures of the drosophila replication program. Genome Res 21: 164–174.

128. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, et al. (2013) Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res 23: 1–11.

129. Mesner LD, Crawford EL, Hamlin JL (2006) Isolating apparently pure libraries of replication origins from complex genomes. Mol Cell 21: 719–726.

130. Woodfine K, Beare DM, Ichimura K, Debernardi S, Mungall AJ, et al. (2005) Replication timing of human chromosome 6. Cell Cycle 4: 172–176.

131. de Moura APS, Retkute R, Hawkins M, Nieduszynski CA (2010) Mathematical modelling of whole chromosome replication. Nucleic Acids Res 38: 5623–5633.

132. Yang SCH, Rhind N, Bechhoefer J (2010) Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. Mol Syst Biol 6: 404.

133. Baker A (2011) Linking the DNA strand asymmetry to the spatio-temporal replication program: from theory to the analysis of genomic and epigenetic data. Ph.D. thesis, Ecole Normale Supérieure de Lyon, Université de Lyon.

134. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660–665.

135. Nicolay S (2006) Analyse des séquences d'ADN par la transformée en ondelettes : extraction d'informations structurelles, dynamiques et fonctionnelles. Ph.D. thesis, University of Liège, Belgium.

136. Zaghloul L (2009) Transcriptional activity, chromatin state and replication timing in domains of compositional skew in the human genome. Ph.D. thesis, Université de Lyon, Ecole Normale Supérieure de Lyon.

137. Alberts B (2007) Molecular biology of the cell. New York: Garland Science.

138. Lodish H (2008) Molecular cell biology. Washington: Macmillan.

139. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature 389: 251–260.

140. Chantalat L, Nicholson JM, Lambert SJ, Reid AJ, Donovan MJ, et al. (2003) Structure of the histone-core octamer in KCl/phosphate crystals at 2.15 A resolution. Acta Crystallogr D Biol Crystallogr 59: 1395–1407.

141. Richmond TJ, Davey CA (2003) The structure of DNA in the nucleosome core. Nature 423: 145–150.

142. Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. Cell 137: 1194–1211.

143. Ptashne M (2007) On the use of the word 'epigenetic'. Curr Biol 17: R233–R236.

144. Abmayr SM, Workman JL (2012) Holding on through DNA replication: histone modification or modifier? Cell 150: 875–877.

145. Chandra T, Kirschner K, Thuret JY, Pope BD, Ryba T, et al. (2012) Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. Mol Cell 47: 203–214.

146. Hyrien O, Goldar A (2010) Mathematical modelling of eukaryotic DNA replication. Chromosome Res 18: 147–161.

147. Baker A, Audit B, Yang SCH, Bechhoefer J, Arneodo A (2012) Inferring where and when replication initiates from genome-wide replication timing data. Phys Rev Lett 108: 268101.

148. van Steensel B, Delrow J, Henikoff S (2001) Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27: 304–308.

149. Park PJ (2009) Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669–680.

150. Schones DE, Zhao K (2008) Genome-wide approaches to studying chromatin modifications. Nat Rev Genet 9: 179–191.

151. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

152. Izenman AJ (2008) Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. New York: Springer.

153. Kaufman L, Rousseeuw PJ (1984) Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons.

154. Julienne H, Zoufir A, Audit B, Arneodo A (2013) Human genome replication proceeds though four chromatin states. PLoS Comput Biol, 9: e1003233.

155. Bickmore WA, van Steensel B (2013) Genome architecture: domain organization of interphase chromosomes. Cell 152: 1270–1284.

156. Rando OJ, Chang HY (2009) Genome-wide views of chromatin structure. Annu Rev Biochem 78: 245-271.

157. Roudier F, Teixeira FK, Colot V (2009) Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. Trends Genet 25: 511–517.

158. Feng S, Jacobsen SE (2011) Epigenetic modifications in plants: an evolutionary perspective. Curr Opin Plant Biol 14: 179–186.

159. The modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330: 1787–1797.

160. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, et al. (2010) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. Nature 471: 480–485.

161. Hon G, Wang W, Ren B (2009) Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol 5: e1000566.

162. Wang Z, Schones DE, Zhao K (2009) Characterization of human epigenomes. Curr Opin Genet Dev 19: 127–134.

163. Lee BK, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, et al. (2012) Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. Genome Res 22: 9–24.

164. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49.

165. Ram O, Goren A, Amit I, Shoresh N, Yosef N, et al. (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell 147: 1628–1639.

166. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.

167. Chessel D, Dufour A, Thioulouse J (2004) The ade4 package -I- One-table methods. R News 4: 5-10.

168. Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20: 53–65.

169. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc: Series B (Stat Methodol) 63: 411–423.

170. Minc E, Courvalin J, Buendia B (2000) Hp1gamma associates with euchromatin and heterochromatin in mammalian nuclei and chromosomes. Cytogenet Cell Genet 90: 279–284.

171. Li Y, Kirschmann DA, Wallrath LL (2002) Does heterochromatin protein 1 always follow code? Proc Natl Acad Sci USA 99 Suppl 4: 16462–16469.

172. Kellum R (2003) HP1 complexes and heterochromatin assembly. Curr Top Microbiol Immunol 274: 53–77.

173. Maison C, Almouzni G (2004) HP1 and the dynamics of heterochromatin maintenance. Nat Rev Mol Cell Biol 5: 296–304.

174. Vakoc CR, Mandat SA, Olenchock BA, Blobel GA (2005) Histone H3 lysine 9 methylation and HP1γ are associated with transcription elongation through mammalian chromatin. Mol Cell 19: 381–391.

175. Smallwood A, Hon GC, Jin F, Henry RE, Espinosa JM, et al. (2012) CBX3 regulates efficient RNA processing genome-wide. Genome Res 22: 1426–1436.

176. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231–1245.

177. Hon G, Wang W, Ren B (2009) Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol 5: e1000566.

178. Tardat M, Murr R, Herceg Z, Sardet C, Julien E (2007) PR-Set7-dependent lysine methylation ensures genome replication and stability through S phase. J Cell Biol 179: 1413–1426.

179. Tardat M, Brustel J, Kirsh O, Lefevbre C, Callanan M, et al. (2010) The histone H4 Lys 20 methyltransferase PR-Set7 regulates replication origins in mammalian cells. Nat Cell Biol 12: 1086–1093.

180. Brustel J, Tardat M, Kirsh O, Grimaud C, Julien E (2011) Coupling mitosis to DNA replication: the emerging role of the histone H4-lysine 20 methyltransferase PR-Set7. Trends Cell Biol 21: 452–460.

181. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA (2007) Identification of higher-order functional domains in the human ENCODE regions. Genome Res 17: 917–927.

182. Škunca N, Altenhoff A, Dessimoz C (2012) Quality of computationally inferred gene ontology annotations. PLoS Comput Biology 8: e1002533.

183. Klein E, Vánky F, Ben-Bassat H, Neumann H, Ralph P, et al. (1976) Properties of the K562 cell line, derived from a patient with chronic myeloid leukemia. Int J Cancer 18: 421–431.

184. Drexler HG (2000) The Leukemia-Lymphoma Cell Line Factsbook. San Diego: Academic Press.

185. Bernardi G (1995) The human genome: organization and evolutionary history. Annu Rev Genet 29: 445–476.

186. Bernardi G (2001) Misunderstandings about isochores. Part 1. Gene 276: 3–13.

187. Eyre-Walker A, Hurst LD (2001) The evolution of isochores. Nat Rev Genet 2: 549–555.

188. Bird AP, Wolffe AP (1999) Methylation-induced repression–belts, braces, and chromatin. Cell 99: 451–454.

189. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465–476.

190. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6–21.

191. Goldar A, Marsolier-Kergoat MC, Hyrien O (2009) Universal temporal profile of replication origin activation in eukaryotes. PLoS One 4: e5899.

192. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188.

193. Julienne H, Zoufir A, Audit B, Arneodo A (2013) Epigenetic regulation of the human genome: coherence between promoter activity and large-scale chromatin environment. Frontiers in Life Science, in press.

194. Ernst J, Kheradpour P, Mikkelsen T, Shoresh N, Ward L, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49.

195. Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R, et al. (2004) A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev 18: 1251–1262.

196. Talasz H, Lindner HH, Sarg B, Helliger W (2005) Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. J Biol Chem 280: 38814–38822.

197. Vakoc CR, Sachdeva MM, Wang H, Blobel GA (2006) Profile of histone lysine methylation across transcribed mammalian chromatin. Mol Cell Biol 26: 9185–9195.

198. Sims JK, Houston SI, Magazinnik T, Rice JC (2006) A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. J Biol Chem 281: 12760–12766.

199. Mackay D (2003) Information Theory, Inference, and Learning Algorithms. Cambridge Univ. Press.

200. Fan JY, Gordon F, Luger K, Hansen JC, Tremethick DJ (2002) The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. Nat Struct Biol 9: 172–176.

201. Tolstorukov MY, Kharchenko PV, Goldman JA, Kingston RE, Park PJ (2009) Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. Genome Res 19: 967–977.

202. Talbert PB, Henikoff S (2010) Histone variants–ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol 11: 264–275.

203. Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton-Carafa Y, et al. (2010) A novel strategy of transcription regulation by intragenic nucleosome ordering. Genome Res 20: 59–67.

204. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.

205. Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA 90: 11995–11999.

206. Ponger L, Duret L, Mouchiroud D (2001) Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res 11: 1854–1860.

207. Antequera F (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60: 1647–1658.

208. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 103: 1412–1417.

209. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39: 457–466.

210. Tang CSM, Epstein RJ (2007) A structural split in the human genome. PLoS One 2: e603.

211. Mohn F, Schübeler D (2009) Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends Genet 25: 129–136.

212. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.

213. Zullo JM, Demarco IA, Pique-Regi R, Gaffney DJ, Epstein CB, et al. (2012) DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell 149: 1474–1487.

214. Green P, Ewing B, Miller W, Thomas PJ, Green ED (2003) Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet 33: 514–517.

215. Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2003) Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett 555: 579–582.

216. Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res 32: 4969–4978.

217. Hiratani I, Gilbert DM (2009) Replication timing as an epigenetic mark. Epigenetics 4: 93–97.

218. Letessier A, Millot GA, Koundrioukoff S, Lachagès AM, Vogt N, et al. (2011) Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature 470: 120–123.

219. De S, Michor F (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat Biotechnol 29: 1103–1108.

220. De S, Michor F (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat Struct Mol Biol 18: 950–955.

221. Ryba T, Battaglia D, Chang BH, Shirley JW, Buckley Q, et al. (2012) Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. Genome Res 22: 1833–1844.

222. Julienne H, Audit B, Arneodo A (2013) Embryonic stem cell specific master replication origins at the heart of the loss of pluripotency. Nucleic Acids Res, submitted.

223. Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. Development 134: 3959–3965.

224. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766–770.

225. Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. Nat Rev Genet 14: 204–220.

226. Bergman Y, Cedar H (2013) DNA methylation dynamics in health and disease. Nat Struct Mol Biol 20: 274–281.

227. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500: 477–481.

228. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, et al. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 6: 479–491.

229. Cremer T, Cremer C (2001) Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cell. Nat Rev Genet 2: 292–301.

230. Gilbert N, Gilchrist S, Bickmore WA (2005) Chromatin organization in the mammalian nucleus. Int Rev Cytol 242: 283–336.

231. Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128: 787–800.

232. Sexton T, Schober H, Fraser P, Gasser SM (2007) Gene regulation through nuclear organization. Nat Struct Mol Biol 14: 1049–1055.

233. Branco MR, Pombo A (2007) Chromosome organization: new facts, new models. Trends Cell Biol 17: 127–134.

234. Fraser P, Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. Nature 447: 413–417.

235. Rando OJ, Chang HY (2009) Genome-wide views of chromatin structure. Annu Rev Biochem 78: 245–271.

236. Meshorer E, Misteli T (2006) Chromatin in pluripotent embryonic stem cells and differentiation. Nat Rev Mol Cell Biol 7: 540–546.

237. Teif VB, Vainshtein Y, Caudron-Herger M, Mallm JP, Marth C, et al. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. Nat Struct Mol Biol 19: 1185–1192.

238. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, et al. (2013) Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell 152: 642–654.

239. Cantone I, Fisher AG (2013) Epigenetic programming and reprogramming during development. Nat Struct Mol Biol 20: 282–289.

240. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, et al. (2013) Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 153: 1149–1163.

241. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947–956.

242. Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, et al. (2006) Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8: 532–538.

243. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315–326.

244. Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G (2007) Genome regulation by polycomb and trithorax proteins. Cell 128: 735–745.

245. Margueron R, Justin N, Ohno K, Sharpe ML, Son J, et al. (2009) Role of the polycomb protein EED in the propagation of repressive histone marks. Nature 461: 762–767.

246. Margueron R, Reinberg D (2011) The Polycomb complex PRC2 and its mark in life. Nature 469: 343–349.

247. Pirotta V, Li HB (2012) A view of nuclear Polycomb bodies. Curr Opin Genet Dev 22: 101–109.

248. Schoenherr CJ, Anderson DJ (1995) The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. Science 267: 1360–1363.

249. Meshorer E, Yellajoshula D, George E, Scambler PJ, Brown DT, et al. (2006) Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Dev Cell 10: 105–116.

250. Bártová E, Galiová G, Krejcf J, Harnicarová A, Strasák L, et al. (2008) Epigenome and chromatin structure in human embryonic stem cells undergoing differentiation. Dev Dyn 237: 3690–3702.

251. Přikrylová T, Pacherník J, Kozubek S, Bártová E (2013) Epigenetics and chromatin plasticity in embryonic stem cells. World J Stem Cells 5: 73-85.

252. Karnani N, Taylor C, Malhotra A, Dutta A (2007) Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. Genome Res 17: 865–876.

253. Weddington N, Stuy A, Hiratani I, Ryba T, Yokochi T, et al. (2008) Replication domain: a visualization tool and comparative database for genome-wide replication timing data. BMC Bioinformatics 9: 530.

254. Ryba T, Battaglia D, Pope BD, Hiratani I, Gilbert DM (2011) Genome-scale analysis of replication timing: from bench to bioinformatics. Nat Protoc 6: 870–895.

255. Simon JA, Kingston RE (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol 10: 697–708.

256. Merkenschlager M, Odom DT (2013) CTCF and cohesin: linking gene regulatory elements with their targets. Cell 152: 1285–1297.

257. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, et al. (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science 303: 343–348.

258. Hou C, Dale R, Dean A (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci USA 107: 3651–3656.

259. Ohlsson R, Lobanenkov V, Klenova E (2010) Does CTCF mediate between nuclear organization and gene expression? Bioessays 32: 37–50.

260. Botta M, Haider S, Leung IXY, Lio P, Mozziconacci J (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Mol Syst Biol 6: 426.

261. Handoko L, Xu H, Li G, Ngan CY, Chew E, et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 43: 630–638.

262. Gaspar-Maia A, Alajem A, Meshorer E, Ramalho-Santos M (2011) Open chromatin in pluripotency and reprogramming. Nat Rev Mol Cell Biol 12: 36–47.

263. Gaspar-Maia A, Alajem A, Polesso F, Sridharan R, Mason MJ, et al. (2009) CHD1 regulates open chromatin and pluripotency of embryonic stem cells. Nature 460: 863–868.

264. Ho L, Crabtree GR (2010) Chromatin remodelling during development. Nature 463: 474–484.

265. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663–676.

266. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131: 861–872.

267. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318: 1917–1920.

268. Jaenisch R, Young R (2008) Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132: 567–582.

269. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.

270. Ruiz S, Panopoulos AD, Herrerías A, Bissig KD, Lutz M, et al. (2011) A high proliferation rate is required for cell reprogramming and maintenance of human embryonic stem cell identity. Curr Biol 21: 45–52.

271. Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, et al. (2008) Global transcription in pluripotent embryonic stem cells. Cell Stem Cell 2: 437–447.

272. Méchali M, Yoshida K, Coulombe P, Pasero P (2013) Genetic and epigenetic determinants of DNA replication origins, position and activation. Curr Opin Genet Dev 23: 124–131.

273. Jones DO, Cowell IG, Singh PB (2000) Mammalian chromodomain proteins: their role in genome organisation and expression. Bioessays 22: 124–137.

274. Nielsen AL, Oulad-Abdelghani M, Ortiz JA, Remboutsika E, Chambon P, et al. (2001) Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. Mol Cell 7: 729–739.

275. Kwon SH, Workman JL (2011) HP1c casts light on dark matter. Cell Cycle 10: 625–630.

276. Tamaru H, Selker EU (2001) A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. Nature 414: 277–283.

277. Jackson JP, Lindroth AM, Cao X, Jacobsen SE (2002) Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature 416: 556–560.

278. Fuks F, Hurd PJ, Deplus R, Kouzarides T (2003) The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res 31: 2305–2312.

279. Lehnertz B, Ueda Y, Derijck AAHA, Braunschweig U, Perez-Burgos L, et al. (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol 13: 1192–1200.

280. Li H, Rauch T, Chen ZX, Szabó PE, Riggs AD, et al. (2006) The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. J Biol Chem 281: 19489–19500.

281. Faust C, Lawson KA, Schork NJ, Thiel B, Magnuson T (1998) The Polycomb-group gene eed is required for normal morphogenetic movements during gastrulation in the mouse embryo. Development 125: 4495–4506.

282. O'Carroll D, Erhardt S, Pagani M, Barton SC, Surani MA, et al. (2001) The polycomb-group gene ezh2 is required for early mouse development. Mol Cell Biol 21: 4330–4336.

283. Peters AH, O'Carroll D, Scherthan H, Mechtler K, Sauer S, et al. (2001) Loss of the suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. Cell 107: 323–337.

284. Tachibana M, Sugimoto K, Nozaki M, Ueda J, Ohta T, et al. (2002) G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. Genes Dev 16: 1779–1791.

285. Dodge JE, Kang YK, Beppu H, Lei H, Li E (2004) Histone H3-K9 methyltransferase ESET is essential for early development. Mol Cell Biol 24: 2478–2486.

286. Pasini D, Bracken AP, Hansen JB, Capillo M, Helin K (2007) The polycomb group protein suz12 is required for embryonic stem cell differentiation. Mol Cell Biol 27: 3769–3779.

287. Vaillant C, Audit B, Arneodo A (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. Phys Rev Lett 99: 218103.

288. Chevereau G, Palmeira L, Thermes C, Arneodo A, Vaillant C (2009) Thermodynamics of intragenic nucleosome ordering. Phys Rev Lett 103: 188103.

289. Milani P, Chevereau G, Vaillant C, Audit B, Haftek-Terreau Z, et al. (2009) Nucleosome positioning by genomic excluding-energy barriers. Proc Natl Acad Sci USA 106: 22257–22262.

290. Chevereau G, Arneodo A, Vaillant C (2011) Influence of the genomic sequence on the primary structure of chromatin. Frontiers in Life Science 5: 29-68.

291. Fudenberg G, Getz G, Meyerson M, Mirny LA (2011) High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat Biotechnol 29: 1109–1113.

292. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

293. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13: 233–245.

294. Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, et al. (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet 7: e1001274.

295. Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural Comput 8: 1341–1390.

296. Jaynes E (2003) Probability Theory The logic of Science. Cambridge Univ. Press.

297. Biesecker LG (2013) Hypothesis-generating research and predictive medicine. Genome Res 23: 1051–1053.

298. Prum B (2010) La démarche statistique. Toulouse: Cépaduès.

299. Prioleau M. private communication.

300. Hyrien O. private communication.

301. Ashoor H, Hérault A, Kamoun A, Radvanyi F, Bajic VB, et al. (2013) HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. Bioinformatics .

302. Lacroix M, Leclercq G (2004) Relevance of breast cancer cell lines as models for breast tumours: fluorescence histograms. J Theor Biol 83: 249–289.